2020

The Statistical Mechanics Of Human Behavior

Christopher William Lynn
University of Pennsylvania

The Statistical Mechanics Of Human Behavior

Abstract

In the study of complex systems, it is often the case that large-scale features emerge from simple properties of the constituent units at the scale below. Nowhere is this observation more evident -- nor are the implications more important -- than in the investigation of human behavior: from the collective firing of thousands or millions of neurons arises the activity of a single brain region, from the communication between of hundreds or thousands of brain regions emerge consciousness and other cognitive functions, and from the interactions between hundreds or thousands of people appear the collective behaviors of human populations. To study such complex systems, cutting-edge research increasingly harkens back to centuries-old insights from statistical mechanics. Here, drawing inspiration from these recent efforts, we adapt and extend methods from statistical mechanics, information theory, and network science to investigate the nature of human behavior across scales.

Generally, the dissertation flows in the direction of decreasing scale, which, coincidentally, approximately corresponds to the chronological order in which the research was produced. We begin in Part I by examining the principles of emergence and control in human populations. In Part II, we study how individual humans learn and process information using networks in the world around them. Finally, in Part III, we investigate whether, and to what extent, the brain operates out of thermodynamic equilibrium. Together, these analyses aim to shed light on the statistical mechanical nature of human behavior.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Physics & Astronomy

First Advisor

Danielle S. Bassett

Keywords

Cognitive Neuroscience, Computational social science, Information theory, Network science, Statistical mechanics

Subject Categories

Biophysics | Physics

THE STATISTICAL MECHANICS OF HUMAN BEHAVIOR

Christopher W. Lynn

A DISSERTATION

in

Physics and Astronomy

Presented to the Faculties of the University of Pennsylvania

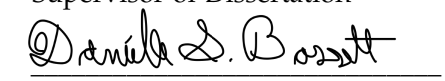
in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

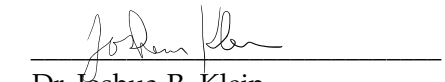
2020

Supervisor of Dissertation



Dr. Danielle S. Bassett
Professor of Bioengineering

Graduate Group Chairperson



Dr. Joshua R. Klein
Professor of Physics and Astronomy

Dissertation Committee

Dr. Daniel D. Lee, Professor of Electrical and Computer Engineering at Cornell Tech
Dr. Randall D. Kamien, Professor of Physics and Astronomy
Dr. Eleni Katifori, Assistant Professor of Physics and Astronomy
Dr. Victor M. Preciado, Associate Professor of Electrical and Systems Engineering

ACKNOWLEDGMENTS

Thank you to all those that have supported me, inspired me, provided a guiding light, helped to expand my mind, and fostered my growth as a scientist and friend. The people to whom I owe a debt of gratitude span not just the last few years, but extend to my earliest days of interest in science. Although I cannot list herein everyone that has helped me along the way, I hope you will consider this page a sincere, albeit far from sufficient, expression of my gratitude.

First, I thank my advisors for patiently supporting me and pushing me to strive for simple answers to insightful questions. In particular, I want to especially thank Prof. Danielle Bassett for providing positive and insightful guidance at every turn, and for opening more doors than I ever thought possible. The genuine compassion that you show for your students does not go unnoticed, and it has been an absolute delight and inspiration having you as my advisor. I also want to thank Prof. Daniel Lee, whose generous mentorship and whiteboard sessions have instilled in me the value of critical and careful reasoning. I also owe a debt of gratitude to Dr. Tanaji Sen and Prof. Dmitrii Makarov, who oversaw my first research experiences and allowed me to pursue my interests, even as I was still figuring out what those were. Finally, to Dr. J for seeing past my high school insolence and inspiring me to study physics. To those listed here, and to the many others that have taken time out of their day to teach me, listen to me, and work with me, I am eternally grateful.

Next, to my committee members, I thank you for graciously directing my research and providing constructive feedback. I specifically want to thank Prof. Randall Kamien, who's vigor for life and knowledge of statistical mechanics never cease to amaze me. Every time we talk I am left with a new question, often about physics. I also want to thank Prof. Victor Preciado, who's generous guidance has helped me to focus and clarify my research interests.

Thirdly, to my friends, you mean the world to me. You have been there when I needed a helping hand, you have provided my with a home away from home, we have shared countless laughs and drinks, and you have each left an imprint on me. In particular, I want to thank Tara, who has quietly inspired me over the last few years. Through your positivity, patience, and empathy, you have supported me along this journey in more ways than you know.

Finally, I want to thank my family. From the hour-long car rides to soccer to the reluctant support of my stylistic choices, your love and encouragement has always shone through. To my parents, I carry with me your countless sacrifices of time and energy. And to my brother, I thank you for helping me see the bigger picture. I am eternally grateful for everything you have done for me.

ABSTRACT

THE STATISTICAL MECHANICS OF HUMAN BEHAVIOR

Christopher W. Lynn

Danielle S. Bassett

In the study of complex systems, it is often the case that large-scale features emerge from simple properties of the constituent units at the scale below. Nowhere is this observation more evident – nor are the implications more important – than in the investigation of human behavior: from the collective firing of thousands or millions of neurons arises the activity of a single brain region, from the communication between of hundreds or thousands of brain regions emerge consciousness and other cognitive functions, and from the interactions between hundreds or thousands of people appear the collective behaviors of human populations. To study such complex systems, cutting-edge research increasingly harkens back to centuries-old insights from statistical mechanics. Here, drawing inspiration from these recent efforts, we adapt and extend methods from statistical mechanics, information theory, and network science to investigate the nature of human behavior across scales.

Generally, the dissertation flows in the direction of decreasing scale, which, coincidentally, approximately corresponds to the chronological order in which the research was produced. We begin in Part I by examining the principles of emergence and control in human populations. In Part II, we study how individual humans learn and process information using networks in the world around them. Finally, in Part III, we investigate whether, and to what extent, the brain operates out of thermodynamic equilibrium. Together, these analyses aim to shed light on the statistical mechanical nature of human behavior.

CONTENTS

LIST OF TABLES

Part I

EMERGENCE AND CONTROL OF COLLECTIVE HUMAN ACTIVITY

In the study of complex systems, it is often assumed that the fundamental building blocks are the pairwise connections between elements. In human populations specifically, pairwise connections combine to form social networks, the early investigations of which marked the birth of network science. But does collective human activity actually emerge from pairwise interactions between individuals? If so, which individuals are most influential in driving the behavior of the population? In Chapter 1, we demonstrate that large-scale patterns in a range of different human activities emerge from simple correlations between pairs of individuals. To do so, we employ maximum entropy modeling techniques, making our description of each population equivalent to an Ising model from statistical mechanics. In Chapters 2-4, we use the Ising model to investigate which nodes (or individuals) in a population are most influential. In Chapter 2, we develop a tractable algorithm for answering this question based on the mean-field approximation. In Chapter 3, we show that the set of influential individuals depends critically on the amount of noise in a system, shifting from central hub nodes in noisy systems to peripheral nodes in deterministic systems. Finally, in Chapter 4, we present a hierarchy of approximation algorithms based on the Plefka expansion that provide increasingly accurate predictions for the set of influential nodes. Together, these results demonstrate that many collective human behaviors can be understood as emerging from networks of pairwise interactions, and that the structure of these interactions determines the optimal strategy for influencing a population.

1

SURGES OF COLLECTIVE HUMAN ACTIVITY EMERGE FROM SIMPLE PAIRWISE CORRELATIONS

This chapter contains work from Lynn, Christopher W., Lia Papadopoulos, Daniel D. Lee, and Danielle S. Bassett. "Surges of collective human activity emerge from simple pairwise correlations." Physical Review X 9.1 (2019): 011022.

Abstract

Human populations exhibit complex behaviors – characterized by long-range correlations and surges in activity – across a range of social, political, and technological contexts. Yet it remains unclear where these collective behaviors come from, or if there even exists a set of unifying principles. Indeed, existing explanations typically rely on context-specific mechanisms, such as traffic jams driven by work schedules or spikes in online traffic induced by significant events. However, analogies with statistical mechanics suggest a more general mechanism: that collective patterns can emerge organically from fine-scale interactions within a population. Here, across four different modes of human activity, we show that the simplest correlations in a population – those between pairs of individuals – can yield accurate quantitative predictions for the large-scale behavior of the entire population. To quantify the minimal consequences of pairwise correlations, we employ the principle of maximum entropy, making our description equivalent to an Ising model whose interactions and external fields are notably calculated from past observations of population activity. In addition to providing accurate quantitative predictions, we show that the topology of learned Ising interactions resembles the network of inter-human communication within a population. Together, these results demonstrate that fine-scale correlations can be used to predict large-scale social behaviors, a perspective that has critical implications for modeling and resource allocation in human populations.

1.1 INTRODUCTION

In the study of human behavior, significant effort has focused on understanding the actions of one or two individuals at a time. It has been observed, for instance, that people engage in "bursts" of actions in quick succession (*49*, *574*, *683*), and significant effort has concentrated on understanding the correlated activity of pairs and triplets of individuals (*198*, *574*). But if we broaden our perspective to an entire population, it becomes increasingly clear that humans also exhibit large-scale patterns of correlated

activity. For example, urban transportation systems undergo surges of correlated activity known as traffic jams (*514*), first responders are required to handle correlated spikes in demand for emergency services (*42*), and internet and telephone networks must be designed to withstand surges of collective activity (*122*, *163*). But where do these large-scale patterns come from? Does it even make sense to discuss such distinct phenomena in the same breath?

Existing explanations for collective human behaviors have focused primarily on external mechanisms, such as fluctuations in urban traffic based on the time of the week (*514*) or spikes in demand for emergency services in response to natural disasters (*42*). While external influences are an important part of the story, such explanations are inherently limited by their reliance on context-specific mechanisms like daily and weekly rhythms and natural disasters. By contrast, interactions between individuals are present in almost every human context, providing the possibility for a much more general explanation for the emergence of large-scale correlations. Precisely this line of reasoning has fostered vibrant efforts linking the study of social systems to tools and intuitions from statistical physics (*126*). By adapting established models of collective behavior in physical systems, such as the Ising model and similar agent-based models, scientists have gained a deeper understanding of the nature of collective behaviors in social systems. This program, for example, has resulted in Ising-like models of social dynamics and human cooperation (*238*, *239*, *516*), viral models aiding in the design of vaccination strategies (*697*), descriptions of the evolution of social networks (*51*), and statistical models of criminal activity (*166*, *303*).

Here we draw inspiration from these seminal results to investigate the role of fine-scale correlations in generating large-scale patterns of human activity. Focusing on four datasets of human activity, from email and private message correspondence to physical contact and music streaming, we find that each population exhibits periods of intense collective activity, which cannot be explained by commonly-used models that assume independence in human behavior (*245*, *290*, *358*, *546*). Intuitively, these surges in activity could be driven by a common external influence, such as people's daily and weekly schedules. Instead, to quantify the collective impact of pairwise correlations, we construct a pairwise maximum entropy model that is formally equivalent to an Ising model from statistical mechanics. While the Ising model has previously been used to understand qualitative aspects of human activity (*126*, *237*, *239*, *240*, *416*), here, in order to make quantitative predictions, we calculate the specific external fields and pairwise interactions that best describe each population. In what follows, we show that this maximum entropy model (i) accurately predicts the frequencies of different patterns of collective human activity, and (ii) bears a close resemblance to the network of inter-human communication within a population. Taken together, these results constitute an important step in the development of quantitative models of collective human behavior based on fine-scale correlations within a population. Such models, in turn, have important implications for resource allocation in communication (*122*) and transportation (*514*) networks, understanding social organization (*490*), and preventing viral epidemics (*513*).

1.2 THE NETWORK EFFECTS OF CORRELATIONS

As a salient example of collective human activity, we begin by studying patterns of email correspondence, focusing specifically on the email activity of 100 scientists at a European research institution over 526 days (*504*, *506*). To understand the role of correlations in the timing of people's actions – and in order to compare against other types of activities that are not directed from one individual to another (*42*, *163*, *185*, *514*, *620*) – we initially focus on the timing of sent emails, while blinding our analysis to the email recipients. Importantly, this will later allow us to compare the architecture of functional interactions derived from our maximum entropy model with the network of communication within the population.

In a sufficiently small window of time Δt, each action appears binary – either individual i sent an email ($\sigma_i = 1$) or they were silent ($\sigma_i = 0$). By discretizing human activity in this way, we can begin to quantify correlations between people's actions. We wish for the time window Δt to be as large as possible (to detect correlations between individuals) without being so large that individuals perform multiple actions within the same window. We find that nearly 90% of consecutive emails from the same individual are sent with at least two minutes in between (Fig. 7.10a), defining a natural time scale that we use as our Δt. Discretizing the data, as shown in Fig. 7.10b, we produce a set of $\sim 3.8 \times 10^5$ binary vectors (patterns) $\boldsymbol{\sigma}$, each of which captures the activity of the entire population within a given two-minute window.

The simplest and most common models of human activity assume that each individual behaves independently, implying that the number of people performing an action in a given window follows a Poisson distribution (*290*). Indeed, the Poisson distribution has been widely used to quantify the effects of various human actions, including telephone calls to a call center (*546*), internet activity (*358*), industrial accidents (*290*, *546*), and highway traffic flow (*245*). In our population of email users, most pairs of individuals are only weakly correlated (Fig. 7.10c), suggesting that small groups should be well-approximated by an independent model. However, if we extend the independent approximation to the entire population of 100 email users, it fails dramatically. While the Poisson distribution predicts a super-exponential drop off in the number of actions performed in a given window, we find instead that human activity actually follows an exponential distribution (Fig. 7.10d). This exponential distribution is characterized by a heavy tail, representing moments in time when many more people are sending emails than would be expected if they were behaving independently. Additionally, we report similar heavy-tailed distributions in separate datasets of private messages, physical contacts, and music streams (Figs. 1.10-1.12). For comparison, after shuffling the timing of the emails to eliminate correlations (*589*), we do not witness a window involving six or more active users (Fig. 7.10d), while we do observe ~1500 such instances in the original dataset – nearly three per day.

The independent approximation also makes straightforward predictions for the rate of each activity pattern. Denoting the probability of individual i sending an email in a given two-minute window by $p_i(\sigma_i)$, the probability of observing a given activity

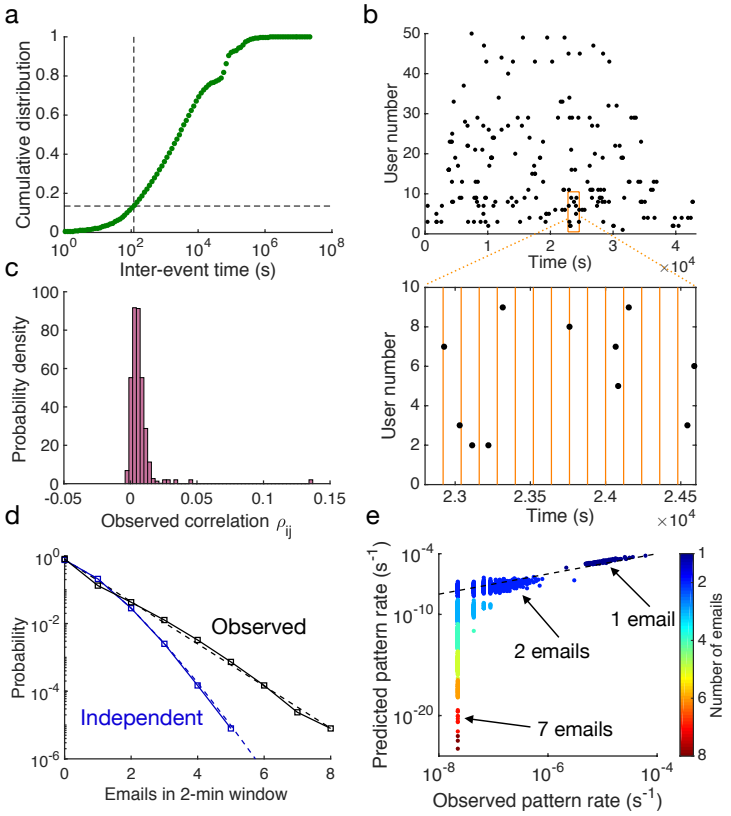


Figure 1.1: Surges of human activity and failure of the independent approximation. (*a*) Distribution of inter-event times for individuals in a network of email correspondence. The dashed lines indicate the proportion of inter-event times less than two minutes. (*b*) Top: Activity of the 50 most active individuals over a half-day period, where each dot represents a sent email. Bottom: Network activity is discretized into two-minute windows. (*c*) Histogram of Pearson correlation coefficients ρ_{ij} between activity time series for all pairs in the 100-person population. (*d*) Distribution of the number of emails sent in a given two-minute window (black) and the distribution after shuffling each person's activity to eliminate correlations (blue). The dashed lines show an exponential distribution fit to the observed data (black) and a Poisson distribution fit to the shuffled data (blue). (*e*) The rate of each observed activity pattern, plotted against the approximate pattern rate assuming independent people. The dashed line indicates equality.

pattern σ is simply predicted to be $P_1(\sigma) = \prod_i p_i(\sigma_i)$. This independent model severely under-predicts patterns involving three or more active email users (Fig. 7.10e), and we find a similar discrepancy in a network of private messages (Fig. 1.10c). In fact, under the independent model, each pattern of email activity involving seven active users should have only appeared roughly once every 10^{20} seconds – longer than the age of the universe. We conclude that the independent approximation fails to explain the heavy-tailed nature of human behavior, characterized by surges of collective activity (*42*, *122*, *163*, *514*). But where do these surges come from?

1.3 A MAXIMUM ENTROPY MODEL OF HUMAN ACTIVITY

To improve upon the independent model, we must take into account correlations between individuals. Intuitively, such correlations could be driven by external influences such as daily and weekly rhythms (Fig. 1.2a), a hypothesis that has dominated existing explanations of large-scale human behaviors (*42*, *122*, *163*, *514*). Alternatively, fine-scale correlations involving only a few individuals could build upon one another to have a strong impact on the population as a whole (Fig. 1.2b). Here, we focus on the simplest possible correlations within a population – those between pairs of individuals – and ask whether these pairwise correlations can give rise to the large-scale patterns of activity that we observe in the data. As we will see, focusing on pairwise correlations represents a natural first step towards understanding emergent collective human activity, opening the door for straightforward generalizations to more complex higher-order correlations (Fig. 1.2b) (*241*, *428*).

We require a model that incorporates the observed pairwise correlations in the data, while including as little information as possible about higher-order correlations between three, four, or more individuals. While it is not immediately obvious how one would construct such a model, Jaynes famously showed that an elegant solution lies in the principle of maximum entropy (*339*): Among the infinite set of distributions consistent with a given set of correlations, the unique one that assumes as little information as possible about additional correlations is precisely the distribution with maximum entropy. This maximum entropy principle lies at the heart of equilibrium statistical mechanics (*161*, *339*) and has become increasingly popular as a tool for studying emergent phenomena in a range of complex systems, including networks of neurons in the brain (*241*, *589*), flocks of birds (*89*), protein structures (*701*), and gene coexpression patterns (*402*). Despite this widespread adoption in biophysics, to our knowledge a similar data-driven approach has not previously been attempted in the social sciences.

Here we consider the pairwise maximum entropy model, defined by the Boltzmann distribution

$$P_2(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_i h_i \sigma_i + \frac{1}{2} \sum_{i \neq j} J_{ij} \sigma_i \sigma_j \right), \tag{1.1}$$

where the external fields h_i and pairwise interactions J_{ij} are Lagrange multipliers that ensure the model matches the observed individual activity rates and pairwise correlations in the data, respectively, and Z is the normalizing partition function. If we switch notation to $\sigma_i = \pm 1$, where $+1$ stands for activity and -1 for inactivity, P_2 is equivalent to the Ising model, which has long been used to simulate human dynamics in social networks (*126*, *237*, *239*, *240*, *416*). However, while existing applications of the Ising model to human populations are based on metaphors about how people interact (*238–240*, *415*, *416*), we emphasize that our use of the Ising model is quantitatively rigorous in the sense that the external fields h_i and interactions J_{ij} are calculated to fit the observed activity of a given population (see Section 1.8.4).

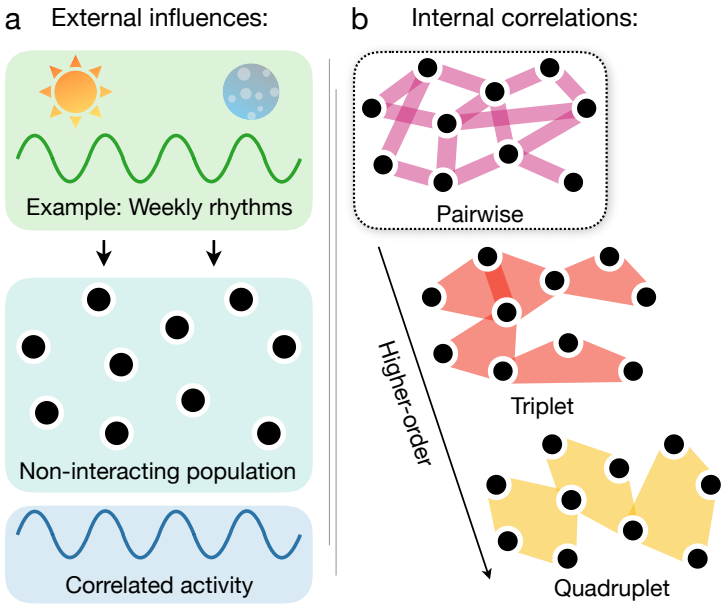


Figure 1.2: External influences versus internal correlations. (*a*) An external mechanism – here taken to be weekly rhythms – influencing the activity of a population of non-interacting humans. Intuitively, circadian and weekly rhythms might influence people to send emails more frequently during the daytime and on weekdays, thereby inducing population-wide correlations. (*b*) Alternatively, population-wide correlations could arise from fine-scale interactions between individuals within a population. The set of all correlations forms a hierarchy, beginning with simple pairwise correlations between two individuals, followed by more complicated higher-order correlations involving three (triplet), four (quadruplet), or more individuals.

1.4 THE MINIMAL CONSEQUENCES OF PAIRWISE CORRELATIONS

Calculations in the Ising model typically require summing over all 2^N activity patterns, where N is the number of elements in a system, prohibiting applications to large populations. Thus, it is common to construct a picture of the whole population by studying many different sub-populations (*589*), such as the 10 email users in Fig. 9.2a. To quantify the explanatory power of pairwise correlations, we need meaningful ways to compare the accuracy of the maximum entropy model P_2 to that of the independent model P_1. Toward this end, we use the Jensen-Shannon divergence $D_{JS}(Q||P)$ as a measure of distance from each of the model distributions (call them Q) to the observed activity distribution P. Put simply, the Jensen-Shannon divergence represents the inverse of the number of independent samples needed to distinguish each model Q from the observed data (*406*). Across 300 random groups of 10 users, we find that on average one would require 3.13×10^4 independent samples – over 43 days worth of data – to distinguish the pairwise model P_2 from the true distribution P (Fig. 9.2b). By contrast, one would typically require five times fewer samples to distinguish the independent model P_1 from the observed data. Moreover, we find qualitatively similar results for individuals engaged in private messaging (Fig. 1.10e), face-to-face interactions (Fig.

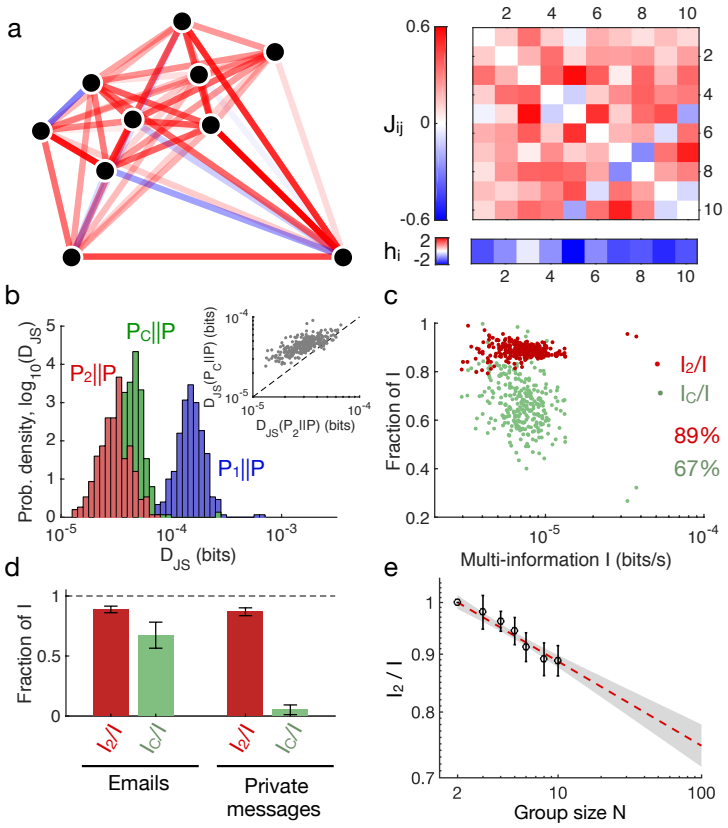


Figure 1.3: The pairwise maximum entropy model accurately describes human behavior. (*a*) Learned Ising interactions J_{ij} and external fields h_i describing a random 10-person group in the email network. (*b*) Jensen-Shannon divergences between the true distribution P and the independent P_1 (blue), maximum entropy P_2 (red), and conditionally independent P_C (green) models. Histograms reflect estimates from 300 random groups of 10 individuals. Inset: $D_{JS}(P_2||P)$ versus $D_{JS}(P_C||P)$ for the 300 groups. The dashed line indicates equality. (*c*) Fraction of the network correlation (quantified by the multi-information I) captured by the maximum entropy (red) and conditionally independent (green) models, plotted against I for each group of 10 people. The multi-information is divided by Δt to remove dependence on the window size. (*d*) Fraction of the total correlation captured by the pairwise (red) and conditionally independent (green) models in four different modes of human activity: email correspondence, private messaging, physical interactions, and online music streaming. Error bars represent standard deviations over 300 random 10-person groups for the email and private message datasets and over 200 groups for the physical contact and music streaming datasets. (*e*) Fraction of the multi-information in the email data captured by the maximum entropy model versus group size, where each data point is averaged over 300 randomly-selected groups. The dashed line represents the best log-linear fit, with 95% confidence interval indicated by the shaded region.

1.11c), and online music streaming (Fig. 1.12c). These observations suggest that the pairwise model provides a marked improvement in accuracy over the independent model.

We also wish to compare against a model representing the hypothesis that patterns of human activity are driven by external influences. While there are many external factors influencing human actions on a daily basis, from weather patterns to shifting demands at work, here we consider the most intuitive and well-studied external influence; namely, the impact of daily and weekly routines (see Fig. 1.2a) (*122*, *163*, *423*, *514*). To formalize the hypothesis that activity patterns are driven by daily and weekly schedules, we consider the conditionally independent model P_C, wherein each individual performs actions independently from all other individuals, but their activity rates are allowed to vary based on the time of the week (*54*, *589*) (see Section 1.8.5). Compared to the conditionally independent model P_C, we find that the maximum entropy model P_2 is closer to the observed data (i.e., has a smaller Jensen-Shannon divergence from P) across 291 of the 300 groups (Fig. 9.2c, Inset). This result is particularly notable when considering that P_2 only has 55 parameters for each group of 10 individuals, while P_C requires knowledge of each individual's email rate at each time during the week, totaling over 5×10^4 parameters.

The pairwise model accurately predicts the rates of particular activity patterns, but does it explain a majority of the total correlation in the population? To answer this question, we note that the total amount of correlation in the network, contributed by correlations between groups of users of all sizes, is quantified by the multi-information $I = S_1 - S$, where S_1 is the entropy of the independent distribution P_1 and S is the entropy of the observed distribution P (*161*) (see Section 1.8.6). To determine the amount of multi-information that is contributed by pairwise correlations, it is useful to review the properties of maximum entropy models. For a population of N elements, we can define a sequence of maximum entropy models P_k that are consistent with all correlations up to the k^{th}-order, where $k = 1, 2, \ldots, N$. These models form a hierarchy, from P_1, in which all elements are independent, up to P_N, which is an exact description of the observed activity. As we climb up this hierarchy, the entropies S_k of the distributions decrease monotonically toward the true entropy ($S_1 \geqslant S_2 \geqslant \cdots \geqslant S_N = S$); and the combined contribution of all k^{th}-order correlations is quantified by the entropy difference $I_k = S_{k-1} - S_k$. We note, for instance, that these entropy differences sum to the full multi-information: $I_2 + \cdots + I_N = I$. Thus, the problem of determining how much of the total correlation in the data stems from pairwise correlations formally reduces to calculating the proportion of the multi-information I that is accounted for by the reduction in entropy from pairwise correlations (i.e., $I_2 = S_1 - S_2$).

We observe that pairwise correlations account for a striking $I_2/I \approx 89\%$ of the total correlation in groups of 10 users (Fig. 9.2c). In turn, this observation implies that the contributions of all other higher-order correlations, $I_3 + \cdots + I_N$, only combine to account for the remaining 11% of the multi-information. Meanwhile, the amount of correlation attributable to daily and weekly rhythms is represented by the entropy difference $I_C = S_1 - S_C$, where S_C is the entropy of the conditionally independent model P_C. This popular explanation for collective human behavior is consistently less effective than the maximum entropy model at capturing the correlations in the data ($I_C/I \approx 67\%$; Fig. 9.2c). Importantly, we show (i) that these results are robust to both

reasonable variation in the time window Δt used to discretize the data (Section 1.8.2.1; Fig. 1.7) as well as differences in the set of individuals selected for analysis (Section 1.8.2.2; Fig. 1.8), and (ii) that the maximum entropy model is relatively consistent over time (Section 1.8.2.3; Fig. 1.9). Moreover, we verify that similar results hold in separate datasets of private messages (Section 1.8.3.1; Fig. 1.10), physical contacts between individuals (Section 1.8.3.2; Fig. 1.11), and music streaming online (Section 1.8.3.3; Fig. 1.12), as summarized in Fig. 9.2d. In the dataset of private messages, for instance, the pairwise model captures nearly the same amount of correlation as in the population of email users ($I_2/I \approx 87\%$), while people's daily and weekly rhythms explain very little of the correlation ($I_C/I \approx 5\%$; Fig. 9.2e). Interestingly, this difference in I/I_C between email activity and private messages (Fig. 9.2c) reflects the commonly-held intuition that email activity is moderately tied to people's work and leisure schedules, while private messages are not.

We are ultimately interested in understanding the role of pairwise correlations in driving large-scale surges of activity in the entire 100-person population. With this goal in mind, we calculate the fraction I_2/I in groups of email users increasing in size from $N = 2$ through 10. For small groups and relatively weak correlations, as the group size increases, we expect the multi-information I to increase in proportion to the entropy difference I_2 (*589*). Indeed, we find that the fraction I_2/I remains nearly constant as the groups grow in size ($I_2/I \propto N^{-0.075\pm0.005}$). Extrapolating to the entire 100-person population, we find with 95% confidence that pairwise correlations account for 72-78% of the total multi-information in the data (Fig. 9.2d). This fraction is especially large when considering the exponential number of possible higher-order correlations ($\sim 2^N$) for populations of increasing size N. We conclude that large-scale patterns of behavior, across several distinct modes of human activity, can be robustly understood as emerging from an underlying network of pairwise correlations.

1.5 MODELING AN ENTIRE POPULATION

Our analysis of relatively small groups indicates that the pairwise maximum entropy model can capture a majority of the correlation structure in groups of up to 100 individuals. This result, in turn, suggests that the heavy-tailed nature of collective human behavior (Fig. 7.10d) – characterized by surges of activity – might emerge organically from pairwise correlations. To test this prediction directly, we must extend the pairwise maximum entropy model to include the entire population of 100 email users. In order to learn the appropriate Ising interactions J_{ij} and external fields h_i for all 100 people, we leverage recent advances in stochastic gradient descent from statistical physics (*215*) and machine learning (*3*), avoiding the exponential complexity of standard Ising calculations (see Section 1.8.4; Fig. 1.13). Fig. 1.4a shows that the pairwise model successfully captures the heavy-tailed nature of human activity, accurately predicting the frequencies of activity surges involving up to seven and eight individuals.

To understand how a network of simple pairwise correlations can generate large-scale spikes in activity, it is useful to study the structure of the Ising parameters in the

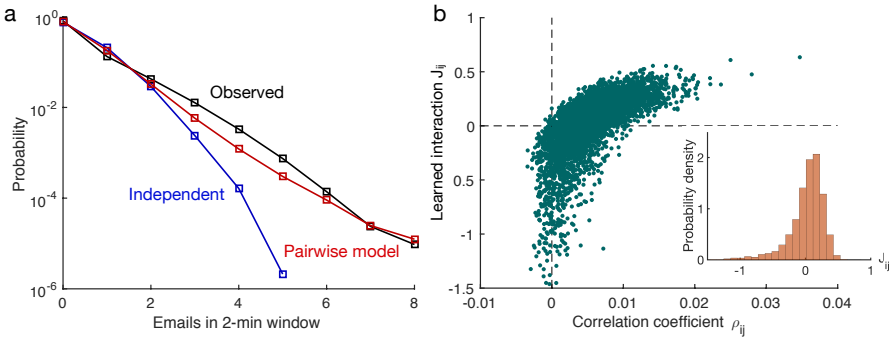


Figure 1.4: Surges of collective activity are captured by pairwise correlations. (*a*) Distribution of the observed number of emails in a given two-minute window (black), the prediction of the independent model (blue), and the prediction of the pairwise maximum entropy model (red). (*b*) Scatter plot illustrating the relationship between the observed pairwise correlations in the data ρ_{ij} and the learned Ising interactions J_{ij} for all pairs in the 100-person population. Inset: Histogram of the learned interactions.

maximum entropy model (Eq. (1.1)). We note that each external field h_i either biases individual i toward activity ($h_i > 0$) or toward inactivity ($h_i < 0$). Meanwhile, each Ising interaction J_{ij} either influences individuals i and j to perform actions at the same time ($J_{ij} > 0$) or at different times ($J_{ij} < 0$). Here, we draw an important distinction between the learned interactions J_{ij} in the maximum entropy model and the observed pairwise correlations ρ_{ij} in the data: while each pairwise correlation quantifies the frequency with which two individuals perform actions at the same time, each Ising interaction represents a functional influence between two individuals to synchronize their activity, thereby inducing a pairwise correlation. Interestingly, while correlations in the network are weak and almost exclusively positive (Fig. 7.10c), the Ising interactions maintain a large amount of heterogeneity (Fig. 1.4b, Inset), with almost an equal number of positive and negative interactions. Indeed, the learned pairwise interactions depend highly non-trivially on the corresponding pairwise correlations in the data (Fig. 1.4b). Importantly, the presence of competing positive and negative interactions generates "frustration," as in spin glasses (*446*), wherein triplets of individuals cannot find a combination of activity and inactivity that simultaneously satisfies all of their interactions. This frustration gives rise to a complex energy landscape of activity patterns with many different local minima, some of which correspond to patterns involving many more active individuals than would be expected under the independent model, thus giving rise to the heavy-tailed behavior in Fig. 1.4a. Intriguingly, such frustrated interactions have previously been hypothesized to drive a number of social phenomena (*126*), such as the formation of coalitions (*236*). By calculating the specific Ising parameters that describe each population, and by identifying the presence of competing positive and negative interactions (Fig. 1.4b, Inset), our work provides rigorous evidence for these long-standing hypotheses.

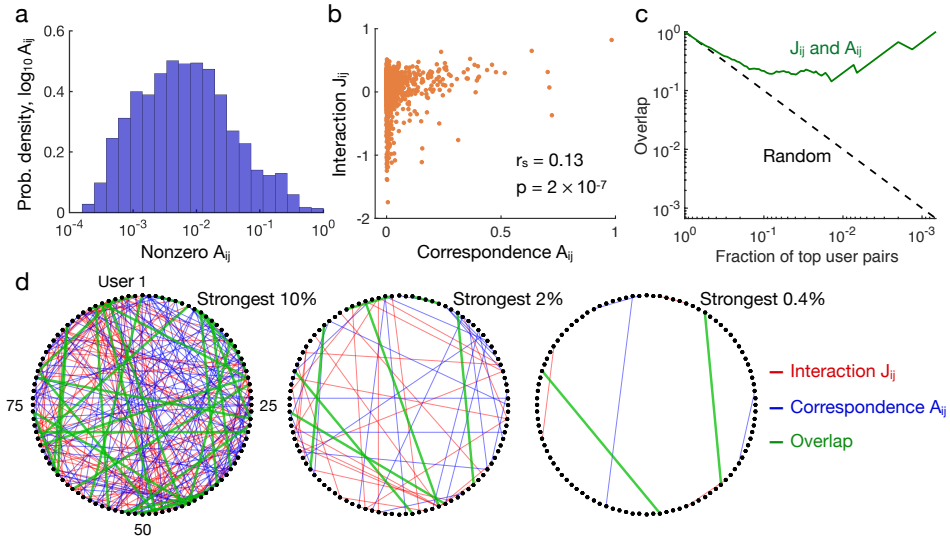


Figure 1.5: The learned pairwise interactions uncover pathways of ground truth communication. (*a*) Histogram of correspondence rates A_{ij} between all pairs of individuals that exchanged at least one email. (*b*) Scatter plot of the learned Ising interactions versus email correspondence rates for pairs that exchanged at least one email. Importantly, J_{ij} and A_{ij} are significantly correlated with Spearman's correlation coefficient $r_s = 0.13$ ($p = 2 \times 10^{-7}$). (*c*) Overlap between the strongest interactions J_{ij} and most frequently corresponding pairs A_{ij} as a function of the fraction of pairs being considered. The dashed line indicates the overlap with a random selection of user pairs. (*d*) Structure of the strongest pairwise interactions (red), highest correspondence rates (blue), and overlap between the two (green) for all 100 individuals. The three networks represent the strongest 10% (left), 2% (middle), and 0.4% (right) of user pairs.

1.6 THE ROLE OF INTER-HUMAN COMMUNICATION

Thus far, we have focused on understanding correlations in the timing of actions, without knowledge of who each person is interacting with in the population. Fundamentally, the Ising interactions J_{ij} are merely learned parameters that ensure consistency with the observed pairwise correlations in the network. However, it is tempting to imbue them with physical significance, interpreting these functional interactions as comprising a network of real-world influences between individuals. For previous applications of maximum entropy models in neuroscience (*241*, *589*) and biology (*89*, *402*, *701*), because comparisons with ground truth interactions are often infeasible, any physical meaning attributed to the learned interactions J_{ij} has remained, at its core, an analogy. By contrast, in the context of email activity, we automatically know a subset of the ground truth interactions – namely, the network of email communication between individuals. Although it is appealing to suspect that the learned Ising interactions are closely related to the structure of email correspondence in the data, we emphasize that this need not be the case. There is an array of circumstances that could influence the activity of two individuals to become correlated, from common functional roles in the network to shared communication with an external third party. Furthermore, even if

correlations do arise from direct communication, this communication could take on many forms that do not appear in the dataset, including face-to-face contact, texts, calls, or other online avenues.

Keeping in mind these reasons for guarded optimism, here we compare the learned interactions J_{ij} from our maximum entropy model with the network of email traffic between individuals. Letting $n_{i\to j}$ denote the number of emails sent from person i to person j, and letting $n_i = \sum_j n_{i\to j}$ denote the total number of emails sent by person i, we define the correspondence rate between two people i and j to be $A_{ij} = (n_{i\to j} + n_{j\to i})/(n_i + n_j)$. In words, A_{ij} represents the fraction of the $n_i + n_j$ emails sent by person i and person j that were addressed to each other. We find that most correspondence only accounts for around 1% of a pair's total email communication, while a small number of pairs communicate almost exclusively with one another (Fig. 1.5a). Considering all pairs of people that exchanged at least one email ($A_{ij} > 0$), we find that the learned Ising interactions J_{ij} are significantly correlated with the correspondence rates A_{ij} in the data (Spearman's correlation coefficient $r_s = 0.13$, $p = 2 \times 10^{-7}$; Fig. 1.5b). This relationship between the learned Ising interactions and the ground truth communication in the population is particularly interesting after reflecting on the myriad ways in which these two networks could have remained unrelated, as described above.

To fully appreciate the strength of the relationship between J_{ij} and A_{ij}, we focus on the fraction f of the strongest pairwise interactions and correspondence rates in the population. These two thresholded networks overlap significantly (Fig. 1.5c), with the strongest 1% of Ising interactions exhibiting a 20% overlap with the top 1% of frequently communicating pairs – 20 times higher than if J_{ij} and A_{ij} were independent. This overlap becomes even more pronounced as we increase the threshold (Fig. 1.5d), such that the single strongest maximum entropy interaction in the entire population corresponds precisely to the pair of individuals that communicate most frequently. This relationship between J_{ij} and A_{ij} provides a compelling mechanistic interpretation for the Ising interactions in our maximum entropy model; namely, frequent communication between a pair of individuals (quantified by A_{ij}) acts as an influence to synchronize their activity (quantified by J_{ij}). As demonstrated in previous sections, the resulting pairwise correlations, in turn, can generate the types of large-scale correlations and surges in human activity that are ubiquitous in the modern world (*42*, *122*, *163*, *185*, *514*, *620*).

1.7 CONCLUSIONS AND FUTURE DIRECTIONS

Despite the widespread investigation of fine-scale correlations as the building blocks of large-scale behavior in complex systems throughout physics (*161*, *339*), neuroscience (*241*, *589*), and biology (*89*, *402*, *701*), a similar quantitative approach to human dynamics has been notably lacking. Here, we provide an important step toward the ultimate goal of understanding the role of fine-scale correlations in generating large-scale patterns of human activity. Studying four datasets that reflect the diversity of human

activity, we first show that all populations exhibit surges of collective activity, a phenomenon that has become the subject of intense research focus (*42*, *122*, *163*, *185*, *514*, *620*). Importantly, these surges in activity cannot be accounted for by commonly-used models that assume independence in human behavior (*245*, *290*, *358*, *546*). To understand where surges in activity come from, we consider the possibility that large-scale patterns arise naturally from combinations of simple pairwise correlations between individuals. To formalize this hypothesis, we utilize the principle of maximum entropy from information theory, deriving a pairwise maximum entropy model of human activity that is formally equivalent to an Ising model. Interestingly, this maximum entropy model accounts for 72-78% of the total correlation in a 100-person population of email users (Fig. 9.2e) and accurately predicts the heavy-tailed distribution of activity surges (Fig. 1.4a). Additionally, we demonstrate that the Ising interactions in our model closely resemble the network of inter-human communication within the population. This close relationship between functional interactions and ground truth communication suggests an intuitive mechanism driving pairwise correlations.

Just as emergent phenomena have garnered significant attention in the natural sciences (*89*, *161*, *241*, *339*, *402*, *428*, *589*, *701*), we anticipate that similar approaches will prove fruitful in the development of accurate models of large social systems. Importantly, while a majority of existing research has focused on the impacts that external influences have on human populations (*42*, *514*), these explanations are fundamentally limited by their reliance on context-specific mechanisms (*122*, *163*). By contrast, interactions between humans are present in almost every context, and, as we have demonstrated, these interactions can build upon one another to have a large-scale impact on the behavior of an entire population. In this way, thinking carefully about the role of fine-scale correlations in activity can have quite general implications for resource allocation in communication (*122*) and transportation (*514*) networks, understanding social organization (*490*), and preventing viral epidemics (*513*).

To conclude, we point out a number of limitations of our analysis that highlight important directions for future work. First, we remark that, given the diversity of experiences that shape human actions, it would be naïve to conclude that all collective behaviors only emerge from internal correlations. To the contrary, it has been well established that external influences play an important role in predicting a number of collective human behaviors (*42*, *122*, *163*, *185*, *514*, *620*). Therefore, future work should investigate the interplay between external influences and internal interactions in human populations. Such an investigation would likely benefit from advances in control theory and influence maximization (*365*, *458*), which have recently been used to predict the propagation of external influences in Ising networks (*415–417*). Second, we note that our investigation has focused primarily on pairwise correlations. While these simplest correlations represent a logical first step, our results do not rule out the possibility that higher-order correlations could also have an important impact on large-scale behavior. Practically speaking, the primary difficulty in studying such higher-order correlations lies in determining which to include in a maximum entropy model, as there exist $\binom{N}{k}$ different choices for each k^{th}-order correlation (a number that grows

nearly exponentially with k). Fortunately, to handle this explosion of parameters, recent advances in neuroscience have produced tractable techniques for generating sparse higher-order maximum entropy models (*241*). Such higher-order models represent systematic generalizations of the methods presented here, and could prove vital for understanding the large-scale impacts of triplet and quadruplet correlations (Fig. 1.2b), which are thought to encode important organizational features in human populations (*198*) (see Section 1.8.7 for an extended discussion).

1.8 SUPPLEMENTARY MATERIAL

1.8.1 *Data preprocessing*

Here, we discuss the details of how the email data is processed, noting that the other datasets follow in an analogous fashion. In total, the dataset contains the email correspondence between 986 members of a European research institution over 526 days (*506*). We focus on the 100 most active individuals, roughly corresponding to the members of the population that sent on average at least one email per day (Fig. 1.6). To quantify correlations between different individuals, we must discretize the data into time bins of width Δt. To choose a suitable bin width, we notice that 90% of consecutive emails from the same individual are sent with at least two minutes in between (Fig. 7.10a), defining a natural time scale that we use as our Δt.Discretizing the 526-day dataset into 2-minute bins, we produce a set of $\sim 3.8 \times 10^5$ binary patterns $\{\boldsymbol{\sigma}\}$ that define the behavior of our population.

1.8.2 *Robustness of the pairwise model*

In Section 1.8.1, we provided first-principles justifications for focusing on the 100 most active individuals in the email dataset and for discretizing the data into bins of width $\Delta t = 2$ minutes. Here, we verify that the success of the pairwise maximum entropy model is robust to reasonable variations in these choices.

1.8.2.1 *Dependence on the bin width*

We investigate the dependence of the pairwise maximum entropy model on the bin width Δt used to discretize the email activity. Throughout, we focus on the 100 most active individuals, and we consider bin widths of $\Delta t = 1$, 5, 10, and 30 minutes. For each value of Δt, we randomly select 200 different groups of 10 individuals and fit

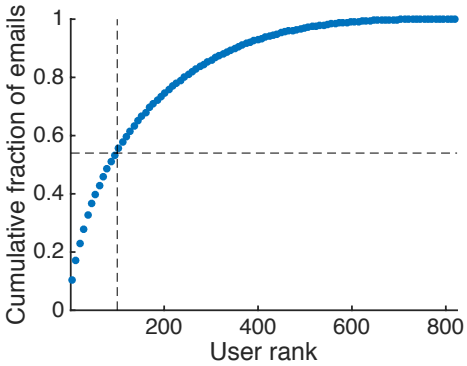


Figure 1.6: Cumulative distribution of emails versus the activity rank of the users. The 100 most active individuals account for 56% of the emails in the network (dashed lines).

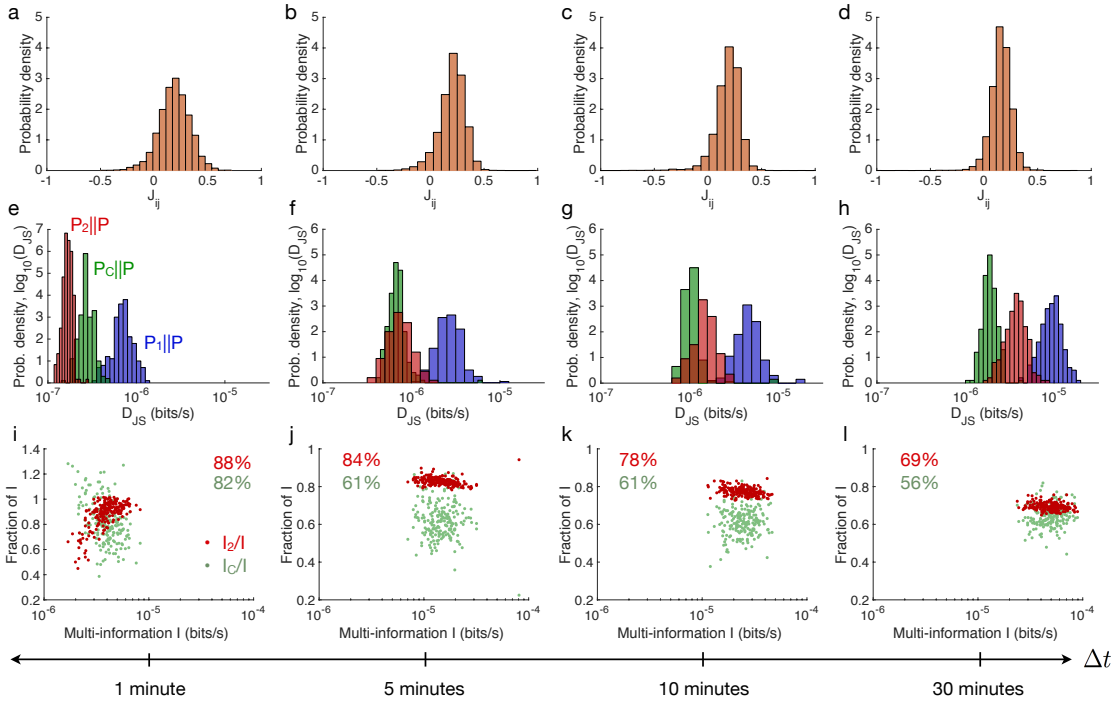


Figure 1.7: Dependence of the pairwise maximum entropy model on the bin width. (*a-d*) Distributions of pairwise couplings for 200 different 10-person groups selected from the 100 most active individuals in the email dataset. From left to right, the data is discretized into bins of width $\Delta t = 1$, 5, 10, and 30 minutes. (*e-h*) Jensen-Shannon divergences between the observed distribution over activity patterns P and the independent P_1 (blue), maximum entropy P_2 (red), and conditionally independent P_C (green) models. The distributions are taken over the 200 groups from panels (*a-d*). (*i-l*) Fraction of the network correlation captured by the maximum entropy (red) and conditionally independent (green) models, plotted against the full network correlation, quantified by the multi-information I. The average percentage of the multi-information captured by each model is displayed in the upper corner. Each dot represents a different group of 10 people, and I is divided by Δt to remove dependence on the window size.

a pairwise maximum entropy model to describe each group. As Δt increases, we witness more windows involving multiple active individuals, thereby strengthening the correlations that we observe in the discretized data. In turn, these stronger correlations give rise to Ising interactions J_{ij} that are more positive and sharply peaked (Fig. 1.7a-d). In Fig. 1.7e-h, we show that the true distribution of activity is approximately five times closer to the maximum entropy model P_2 than to the independent model P_1 across all values of Δt considered, demonstrating the consistency of the pairwise model in predicting human behavior. On the other hand, the performance of the conditionally independent model P_C increases significantly as Δt increases, even outperforming the pairwise model for $\Delta t \geqslant 10$ minutes. We note, however, that for such large bin widths, people often send multiple emails within the same window, and treating the data as binary may not be justified. In Fig. 1.7i-l, we see that the pairwise model captures

nearly all of the multi-information in the 10-person groups across all choices for Δt. By contrast, the conditionally independent model consistently captures a smaller fraction of the multi-information in the data. Furthermore, for $\Delta t = 1$ minute, the conditionally independent model has lower entropy than the data itself (i.e., $I_C/I > 1$) for 30 of the 200 groups, which is a clear indication that the model is overfitting the data.

1.8.2.2 *Dependence on the individuals being analyzed*

We investigate the dependence of the maximum entropy model on the set of individuals chosen for analysis. In particular, we consider 200 different 10-person groups selected from among the 100 most active email users, the 400 most active users, and all 824 users that sent at least one email. Throughout this section, the bin width is fixed at $\Delta t = 5$ minutes. As we focus on more active individuals, the observed correlations become stronger, which is reflected in the fact that the distribution of learned interactions J_{ij} among the top 100 individuals is more sharply peaked and positive than the pairwise interactions between the top 400 and all 824 individuals (Fig. 1.8a-c). In Fig. 1.8d-f, we again find that the pairwise model is approximately five times closer to the true distribution than the independent model across all three subpopulations. By contrast, the conditionally independent model performs nearly as well as the pairwise model among the 100 most active individuals, but provides only marginal improvements over the independent model for all 824 individuals. The failure of the conditionally independent model in describing the entire 824-person population is not surprising given that most individuals sent less than one email every five days, leaving daily and weekly rhythms with little to no predictive power.

We now consider the fraction of the multi-information captured by each model. For all 824 individuals, Fig. 1.8g shows that the conditionally independent model captures a slightly larger fraction of the multi-information than the maximum entropy model; however, P_C erroneously includes more correlation than the data itself (i.e., $I_C/I > 1$) for 20 of the 200 groups of 10 people, indicating that the model is overfitting the data. For both the top 100 and 400 most active individuals, the maximum entropy model captures a significantly larger fraction of the network correlation than the conditionally independent model (Fig. 1.8h-i). We conclude that the predictions of the pairwise maximum entropy model are robust to variations in both the bin width Δt as well as the set of individuals chosen for analysis.

1.8.2.3 *Consistency of the pairwise model over time*

By employing the pairwise maximum entropy model in Eq. (1.1), we implicitly assume that the population activity can be modeled as a stationary distribution; that is, that the local fields h_i and interactions J_{ij} do not change over time. Here, we test this assumption explicitly while noting that the development of time-evolving maximum entropy models is an important direction for future work (see Section 1.8.7.3 for an extended discussion). Specifically, we wish to determine if the Ising parameters describing one portion of the email activity resemble those describing another portion

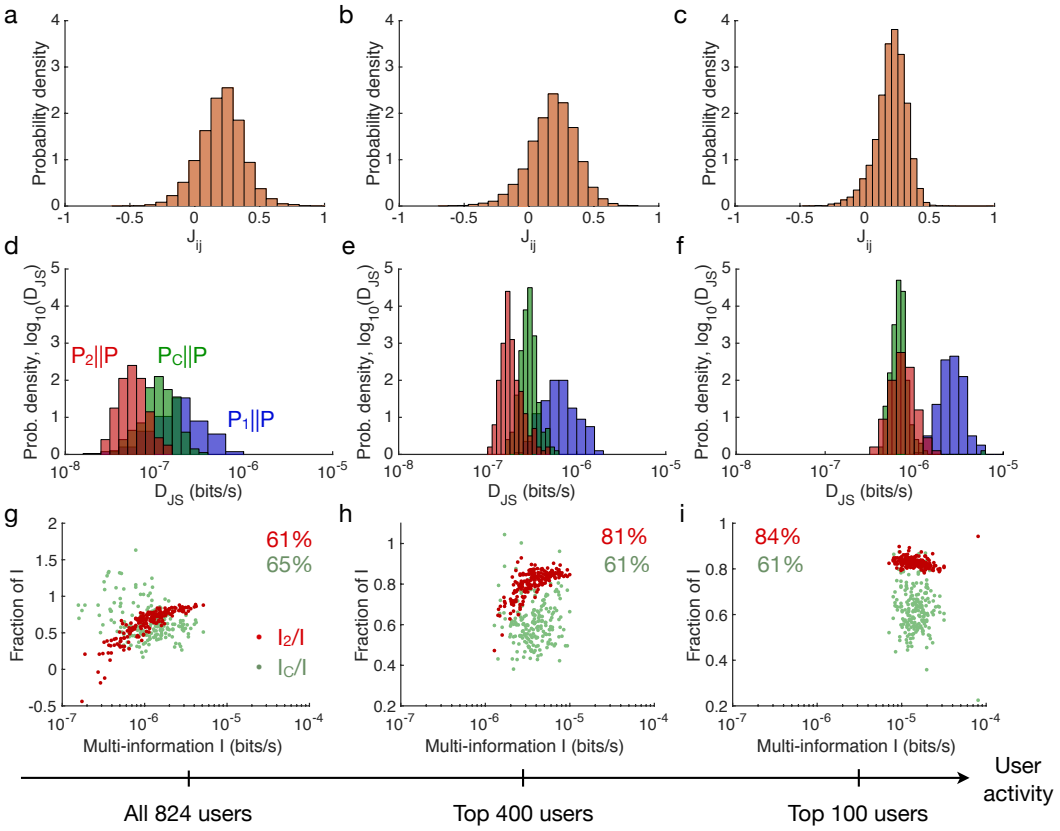


Figure 1.8: Dependence of the pairwise model on the set of individuals chosen for analysis in the email dataset. (*a-c*) Distributions of pairwise interactions for 200 different groups of 10 individuals, where the data is discretized with bin width $\Delta t = 5$ minutes. From left to right, the 200 groups are chosen from among all 824 people that sent at least one email, the 400 most active individuals, and the 100 most active individuals, respectively. (*d-f*) Jensen-Shannon divergences between the observed distribution over activity patterns P and the independent P_1 (blue), maximum entropy P_2 (red), and conditionally independent P_C (green) models. The distributions are taken over the 200 groups of users. (*g-i*) Fraction of the network correlation captured by the pairwise maximum entropy (red) and conditionally independent (green) models, plotted against the full network correlation, quantified by the multi-information I. The average percentage of the multi-information captured by each model is displayed in the upper corner. The multi-information is divided by Δt to remove dependence on the window size.

of the activity. To do so, we divide the dataset into two halves corresponding roughly to the first and last 263 days of email activity. Fig. 1.9a-c shows that the statistics describing the population activity remain remarkably consistent over time, with both the user activity rates and pair correspondence rates A_{ij} being strongly correlated between the two halves of data (Pearson's correlations $r_p = 0.77$ for the activity rates and $r_p = 0.91$ for the correspondence rates).

To study the consistency of the maximum entropy model, we randomly select 200 different 10-person groups from among the 74 users that sent at least one email in both halves of the dataset, and we then learn pairwise models describing each group for

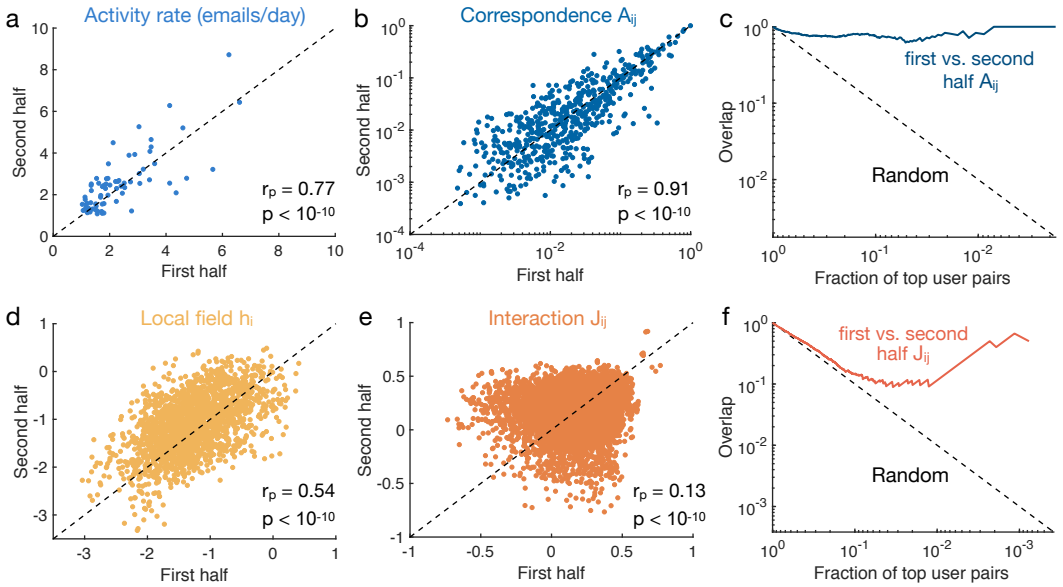


Figure 1.9: Consistency of the pairwise maximum entropy model over time. (*a*) Comparison of email user activity rates in the first half versus the second half of the dataset; the dashed line indicates equality. (*b*) Correspondence rates A_{ij} between pairs of users are strongly correlated across the two halves of the dataset. (*c*) Overlap between the most frequently corresponding pairs of users in the first half and those in the second half as a function of the fraction of pairs being considered. The dashed line indicates the overlap with a random selection of user pairs. (*d*) For 200 random groups of 10 individuals, we compare the local fields h_i of pairwise maximum entropy models fit to either the first or second half of the email data. (*e*) For the same 200 random groups, we compare the Ising interactions J_{ij} of the pairwise models fit to the two halves of the dataset. (*f*) For each half of the dataset, we average the interactions J_{ij} over all 200 groups and plot the overlap between average interaction networks as a function of the fraction of user pairs being considered. As in panel (*c*), the dashed line indicates the overlap with a random selection of pairs.

each half of data. Fig. 1.9d-e shows that the local fields h_i and interactions J_{ij} modeling the population activity are significantly correlated over time (Pearson's correlations $r_p = 0.54$ for the local fields and $r_p = 0.13$ for the interactions). The consistency of the Ising interactions J_{ij} between the two halves of data becomes even more apparent when we focus on the strongest interactions in the population (Fig. 1.9f). Together, these results indicate that the patterns of population activity remain relatively consistent over time, justifying our application of the stationary maximum entropy model as a first step toward more complex dynamical models.

1.8.3 *Other modes of human activity*

In the main text, our analysis focused primarily on a dataset of email activity. Here, we independently verify the ability of the pairwise maximum entropy model to quantita-

tively describe collective human behavior in three other datasets representing a diverse range of human activities.

1.8.3.1 *Private messages*

We first consider a dataset of $\sim 6 \times 10^5$ private messages sent between 1899 students at U.C. Irvine over the span of 193 days (*504*). As in the context of email activity, we focus on the individuals that sent on average at least one message per day, corresponding to the 66 most active students in the population. To choose an appropriate bin width, we consider the distribution of time gaps between consecutive messages from the same student (Fig. 1.10a). Comparing against the equivalent distribution in the email dataset (Fig. 1.6b), we notice that many more private messages than emails are sent with short gaps ($\lesssim$ 1 minute) in between. This bursty behavior indicates that the private messages serve as a more conversational communication medium than emails, a fact that will later help in understanding the impact of daily and weekly rhythms. Due to the bursty nature of private messages, we reduce our bin width to $\Delta t = 1$ minute, yielding a dataset of $\sim 2.8 \times 10^5$ binary activity patterns.

As in the network of email correspondence, the independent model P_1 fails to explain the collective behavior in the private message population (*245*, *290*, *358*, *546*); while the independent model predicts a super-exponential drop off in the number of active individuals in a given window, we find that the distribution of private messages is actually heavy-tailed, fitting closely to an exponential distribution (Fig. 1.10b). Additionally, in Fig. 1.10c we see that the independent model dramatically under-predicts patterns involving two or more active individuals. To improve upon the independent model, we again consider two competing hypotheses: (i) that large-scale patterns emerge from an aggregation of simple pairwise correlations (represented by the pairwise maximum entropy model P_2), and (ii) that large-scale patterns are driven by similarities in people's weekly routines (represented by the conditionally independent model P_C). Randomly selecting 300 groups of 10 people, Fig. 1.10d shows that the pattern rates predicted by the pairwise maximum entropy model are tightly correlated with the observed pattern rates, avoiding the inaccuracies of the independent and conditionally independent models.

Additionally, calculating the Jensen-Shannon divergences $D_{JS}(Q||P)$ from each model Q to the observed data P, we find that one would typically need over five times more samples to distinguish the pairwise model than the independent model (Fig. 1.10e), reflecting roughly the same performance as in the network of email correspondence. Interestingly, in contrast to email activity, the conditionally independent model provides nearly no improvement over the independent model in the dataset of private messages. Additionally, Fig. 1.10f shows that the pairwise maximum entropy model captures $I_2/I \approx 87\%$ of the correlation in the data, nearly identical to its performance on the network of email correspondence, while the conditionally independent model accounts for a strikingly small fraction of the correlation structure ($I_C/I \approx 5\%$). This difference in the performance of P_C between the private message and email datasets

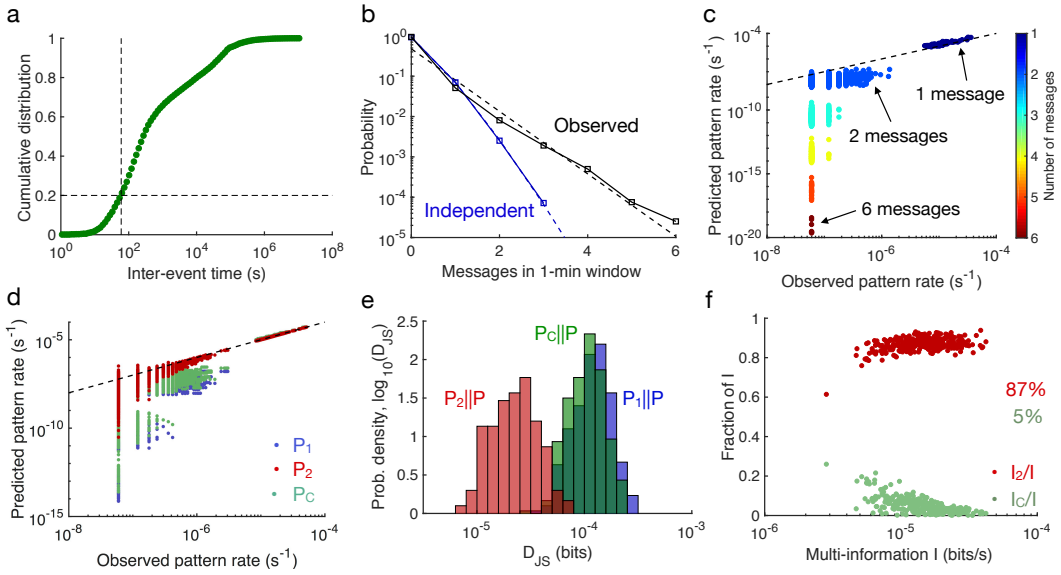


Figure 1.10: Performance of the pairwise maximum entropy model in a dataset of private messages. (*a*) Cumulative distribution of inter-event times for the 66 most active individuals. Approximately 80% of consecutive messages from the same person are sent with at least one minute in between (dashed lines). (*b*) Distribution of the messages sent in a given one-minute window in the dataset (black) and after shuffling individuals' activities to eliminate correlations (blue); dashed lines indicate an exponential fit to the observed data (black) and a Poisson fit to the shuffled data (blue). (*c*) The rate of each observed activity pattern, plotted against the approximate rate under the independent model P_1; the dashed line indicates equality. (*d*) We plot the rate of each observed activity pattern across 300 randomly selected groups of 10 individuals against the approximate rates under the independent model P_1 (blue), the pairwise maximum entropy model P_2 (red), and the conditionally independent model P_C (green); the dashed line indicates equality. (*e*) Jensen-Shannon divergences between the true distribution P and the independent P_1 (blue), maximum entropy P_2 (red), and conditionally independent P_C (green) models; the histograms reflect estimates from the 300 10-person groups. (*f*) Fraction of the network correlation (i.e., multi-information I) captured by the pairwise (red) and conditionally independent (green) models, plotted against the full multi-information. We note that I is divided by Δt to remove the dependence on window size.

suggests that the conversational nature of private messages makes them less likely than email traffic to depend on people's routines. By contrast, the maximum entropy model accurately describes the activity in both populations, further validating the conclusion that patterns of collective behavior can be understood as emerging from simple pairwise correlations.

1.8.3.2 *Physical contacts*

Thus far, we have only studied human actions mediated by online communication. Here, we instead consider a dataset of face-to-face interactions between 50 attendees at the ACM Hypertext 2009 conference, which spanned three days (*331*). Discretizing the population activity into bins of width $\Delta t = 20$ seconds, we arrive at a set of $\sim 10^4$ binary

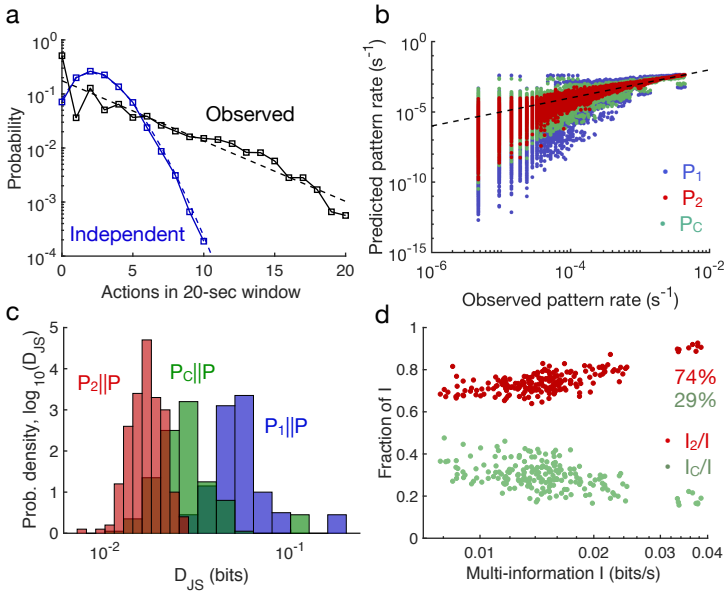


Figure 1.11: Performance of the pairwise model in a dataset of face-to-face contacts between individuals. (*a*) Distribution of the number of contacts in a given 20-second window observed in the dataset (black) and after shuffling individuals' activities to eliminate correlations (blue); dashed lines indicate an exponential fit to the observed data (black) and a Poisson fit to the shuffled data (blue). (*b*) The rate of each observed activity pattern across 200 randomly selected groups of 10 individuals is plotted against the approximate rates under the independent model P_1 (blue), the pairwise maximum entropy model P_2 (red), and the conditionally independent model P_C (green); the dashed line indicates equality. (*c*) Jensen-Shannon divergences between the true distribution P and the independent P_1 (blue), maximum entropy P_2 (red), and conditionally independent P_C (green) models; the histograms reflect estimates from the 200 10-person groups. (*d*) Fraction of the network correlation (i.e., multi-information I) captured by the pairwise (red) and conditionally independent (green) models, plotted against the full multi-information; I is divided by $\Delta t = 20$ seconds to remove the dependence on window size.

activity vectors. As in both the networks of email and private message correspondence, we observe that the number of human contacts within a given 20-second window roughly obeys an exponential distribution, while the independent model instead predicts a Poisson distribution that severely under-predicts the likelihood of surges in human activity (Fig. 1.11a). To study the pairwise maximum entropy model, we generate 200 random groups of 10 individuals. Fig. 1.11b shows that the rates of activity patterns predicted by the pairwise model are tightly correlated with the rates at which they were observed at the conference, providing consistently more accurate predictions than both the independent and conditionally independent models.

Quantitatively, one would require three to four times as many samples to distinguish the independent model from the observed data than the maximum entropy model, and the maximum entropy model achieves a lower Jensen-Shannon divergence from the observed data than the conditionally independent model across all 200 groups of attendees (Fig. 1.11c). Additionally, Fig. 1.11d shows that the pairwise model captures

$I_2/I \approx 74\%$ of the correlation in the face-to-face contacts. While this is slightly lower than that observed for emails and private messages, we remark that the conditionally independent model only accounts for $I_C/I \approx 29\%$ of the correlation in the data. Interestingly, despite physical interactions representing a quite different mode of human activity from online communication, we still find that patterns of population behavior are well-described as arising from pairwise correlations.

1.8.3.3 *Music streams*

To this point, all of our analysis has focused on modes of human activity that are themselves types of interactions between individuals. It is natural to suspect, therefore, that these activities might be particularly conducive to being described by a pairwise model. To test the ability of the pairwise maximum entropy model to describe other modes of human activity, here we consider a dataset of 610 individuals streaming music on the website `last.fm` over the span of one year (*131*). Discretizing the streaming activity into bins of width $\Delta t = 150$ seconds (roughly corresponding to the length of an average song), we arrive at a set of $\sim 2 \times 10^5$ activity vectors. Considering the number of music streams in a given 150-second window, we notice that the observed distribution is notably not described by an exponential distribution (Fig. 1.12a), which is attributable to the fact that the streaming data is much less sparse than any of the three activities studied previously. Nevertheless, we still find that the observed distribution is heavy-tailed relative to the independent Poisson distribution, and is characterized by surges of activity where upwards of 50 users are streaming music at a given time.

Randomly selecting 200 groups of 10 users, we show in Fig. 1.12b that the pairwise maximum entropy model provides a much tighter fit of the observed activity pattern rates than either the independent or conditionally independent models. Moreover, by studying the Jensen-Shannon divergences between the different models and the observed distribution of activity patterns, we find that we would need over six times as many data samples to distinguish P_2 from P than to distinguish P_1 from P and over four times more samples to distinguish P_C (Fig. 1.12c). These results are further supported by Fig. 1.12d, which shows that the pairwise model captures $I_2/I \approx 74\%$ of the correlation in groups of 10 users, nearly identical to the case of face-to-face contacts. Meanwhile, the daily and weekly rhythms only account for $I_C/I_N \approx 35\%$ of the correlation in the data.

All together, our analysis of private messages, face-to-face contacts, and online music streams serve to strengthen the conclusions made in the main text; namely, that pairwise correlations can build upon one another to generate predictable patterns of population-wide activity.

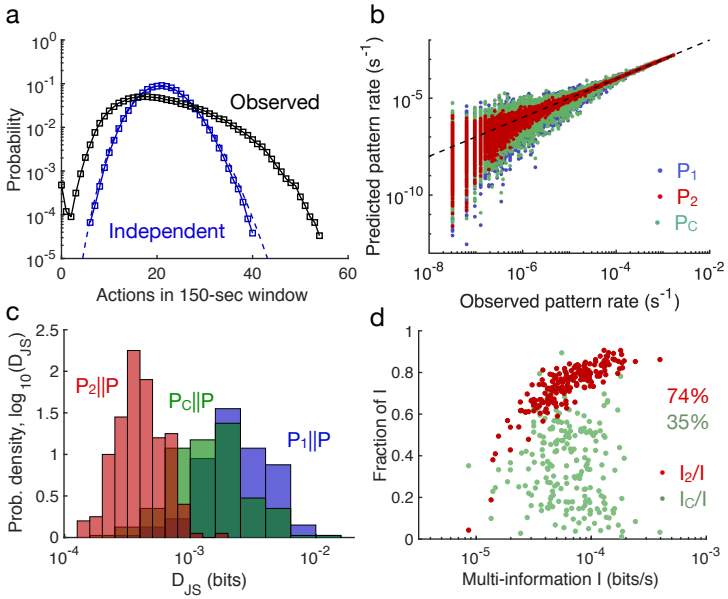


Figure 1.12: Performance of the maximum entropy model in a dataset of music streams. (*a*) Distribution of the number of streams in a given 150-second window in the dataset (black) and after shuffling individuals' activities to eliminate correlations (blue); dashed line indicates a Poisson fit to the shuffled data (blue). (*b*) The rate of each observed activity pattern across 200 randomly selected groups of 10 individuals is plotted against the approximate rates under the independent model P_1 (blue), the pairwise maximum entropy model P_2 (red), and the conditionally independent model P_C (green); the dashed line indicates equality. (*c*) Jensen-Shannon divergences between the true distribution P and the independent P_1 (blue), maximum entropy P_2 (red), and conditionally independent P_C (green) models; the histograms reflect estimates from the 200 10-person groups. (*d*) Fraction of the network correlation (i.e., multi-information I) captured by the pairwise (red) and conditionally independent (green) models, plotted against the full multi-information; I is divided by $\Delta t = 150$ seconds to remove the dependence on window size.

1.8.4 *Learning a pairwise maximum entropy model: The inverse Ising problem*

Here we present the theory and methodology behind learning a pairwise maximum entropy model of collective human activity. Specifically, we describe how to calculate the Ising parameters h_i and J_{ij} from a dataset of collective activity patterns. This inference task has a rich history in machine learning under the title Boltzmann machine learning (3) and is commonly referred to in physics as the inverse Ising problem (33).

1.8.4.1 *Exact models for small populations*

Given the observed distribution P of activity patterns, there is a unique pairwise model P_2 that is consistent with the observed activity rates $\langle \sigma_i \rangle$ and pairwise correlations $\langle \sigma_i \sigma_j \rangle$, where $\langle \cdot \rangle$ represents an average over P. To learn this pairwise model, one

typically begins with an initial pairwise distribution Q with parameters $\tilde{h}_i$ and $\tilde{J}_{ij}$ and then performs gradient descent in the model parameters, with gradients defined by

$$\Delta\tilde{h}_i \propto \langle\sigma_i\rangle - \langle\sigma_i\rangle_Q \,, \tag{1.2}$$

$$\Delta\tilde{J}_{ij} \propto \langle\sigma_i\sigma_j\rangle - \langle\sigma_i\sigma_j\rangle_Q \,, \tag{1.3}$$

where $\langle\cdot\rangle_Q$ represents an average over Q. For groups of size $N = 10$, these gradient calculations are tractable and standard gradient descent converges to the correct pairwise maximum entropy model P_2.

1.8.4.2 *Approximate models for large populations*

The primary difficulty in learning a maximum entropy model for a large population, such as the group of 100 email users, lies in calculating the one- and two-point correlations under Q at each gradient step in Eqs. (1.2) and (1.3). For large populations, exact calculations using the Boltzmann distribution are infeasible, and one must resort to approximate methods. The standard strategy is to simulate the system using Monte Carlo techniques (*241*, *249*, *665*). Naïvely, one would run a new Monte Carlo simulation to estimate the gradients at each step of the learning algorithm. However, this straightforward approach is extremely inefficient. Instead, one can adjust the estimates of the one- and two-point correlations at each gradient step using importance sampling (*347*) or histogram Monte Carlo (*215*). In addition to limiting the number of Monte Carlo simulations, because each sample $\boldsymbol{\sigma}$ of Q is dominated by inactive individuals, one can leverage sparse matrix operations to significantly speed up the simulations themselves.

We terminate the learning algorithm when the model correlations, $\langle\sigma_i\rangle_Q$ and $\langle\sigma_i\sigma_j\rangle_Q$, are sufficiently close to the observed correlations. The relevant scale for errors in the observed correlations is defined by the standard deviations $\Delta\langle\sigma_i\rangle$ and $\Delta\langle\sigma_i\sigma_j\rangle$, which are estimated by bootstrap sampling from the original dataset. Thus, the learning algorithm is terminated when

$$|\langle\sigma_i\rangle - \langle\sigma_i\rangle_Q| < \Delta\langle\sigma_i\rangle \approx 2.2\times10^{-4} \tag{1.4}$$

$$|\langle\sigma_i\sigma_j\rangle - \langle\sigma_i\sigma_j\rangle_Q| < \Delta\langle\sigma_i\sigma_j\rangle \approx 1.7\times10^{-4}. \tag{1.5}$$

We confirm that the individual email rates and pairwise correlations under the maximum entropy model P_2 match the observed correlations within the experimental errors in the data (Fig. 1.13a-c).

For a population of 100 individuals, defining a pairwise maximum entropy model requires learning $N(N+1)/2 = 5050$ different parameters. Given such a large number, it is possible that the model is being finely tuned to match statistical errors in the data. To test for overfitting, we randomly select 476 of the 526 days to learn the model, and then we test the accuracy of the model on the remaining 50 days. We confirm that the pairwise model assigns the same amount of probability to the test data as to the training data, within errors, demonstrating that the learned model generalizes

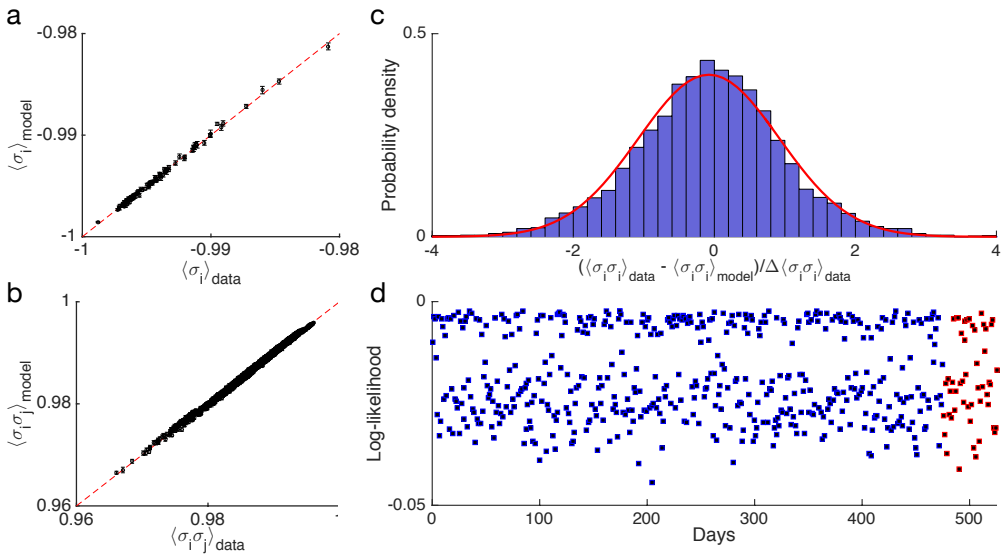


Figure 1.13: Learning a pairwise maximum entropy model for a 100-person population. (*a*) Reconstructed activity rates for all 100 individuals under the maximum entropy model, plotted against their true activity rates. The dashed line indicates equality. (*b*) Reconstructed pairwise correlations under the maximum entropy model versus the observed correlations. (*c*) Distribution of the differences between the true and model pairwise correlations, normalized by the error in the data $\Delta\langle\sigma_i\sigma_j\rangle$. For reference, the red line is a Gaussian distribution with unit variance. The empirically measured distribution has nearly Gaussian shape with standard deviation ≈ 1.05, demonstrating that the learning algorithm reconstructs the pairwise correlations within experimental precision. (*d*) The per-person average log-likelihood of the data $\langle \log P_2(\boldsymbol{\sigma})\rangle / N$, where the average is taken over all patterns within a given day, computed for the training days (blue) and test days (red). The data has been sorted so that the test days follow the training days, but the true choice of test days was random.

to describe data outside of the training set (Fig. 1.13d). We conclude that the learned pairwise model (i) fits the activity data within experimental precision and (ii) does not overfit statistical noise in the data. For access to the calculated external fields h_i and pairwise interactions J_{ij}, please contact the corresponding author.

1.8.5 *The conditionally independent model*

To test the prediction that collective behavior is driven by similarities in people's daily and weekly routines, we study the conditionally independent model P_C. Letting $p_i^t(\sigma_i)$ denote the probability of person i performing an action within a window of width Δt at time t during the week, the conditionally independent model is defined by

$$P_C(\boldsymbol{\sigma}) = \frac{\Delta t}{\omega} \sum_t \prod_i p_i^t(\sigma_i), \tag{1.6}$$

where ω denotes the length of a day or week. Under this conditionally independent model, correlations between individuals are driven by fluctuations in their inherent activity rates.

1.8.6 *Estimating entropy from finite data*

To calculate the multi-information $I = S_1 - S$ of the network activity, we must compute the entropies of the independent model S_1 and the observed data S. While calculating S_1 is straightforward, we must estimate the true entropy S from a finite number of samples, possibly leading to finite-size errors. Suppose that the dataset consists of the patterns $\{\sigma^\alpha\}$ with corresponding probabilities $\{p^\alpha\}$. One could naïvely estimate the entropy using the standard formula

$$\tilde{S} = -\sum_\alpha p^\alpha \log p^\alpha. \tag{1.7}$$

However, since some of the patterns are likely missing and the probabilities p^α are not exact, this estimate should fundamentally be viewed as an approximation to S that improves as the number of samples increases. To correct for the sample size dependence of $\tilde{S}$, we sub-sample the data and fit the resulting estimates using a form proposed by Strong et al. (*640*),

$$\tilde{S}(\text{size}) = S + \frac{a}{\text{size}} + \frac{b}{\text{size}^2}, \tag{1.8}$$

where a and b are finite-size corrections. Using this fit, we can extract an accurate estimate of the true entropy S. We remark that for large datasets such as those considered here, and for relatively small networks like the 10-person groups studied in the main text, finite-size errors are small.

1.8.7 *Extended discussion*

Our investigation of collective human behavior yields three distinct conclusions:

1. Large-scale behavior, characterized by surges in collective activity, cannot be understood using models that assume humans behave independently.
2. While collective behavior is far from independent, the minimal extension of the independent model consistent with the observed pairwise correlations captures most of the correlation in all populations considered, accurately predicting surges of collective activity.
3. In the network of email correspondence, the learned pairwise interactions are closely related to the underlying topology of inter-human communication, imbuing the maximum entropy model with real-world interpretability.

Here we discuss the implications and limitations of these results, while keeping in mind that modern life involves a diverse range of activities, some of which may require a

fundamentally different approach. Throughout, we emphasize important opportunities for future research.

1.8.7.1 *Internal correlations versus external influences*

In the study of human dynamics, as in the study of physical and biological systems, any macroscopic behavior that evades explanation by a model of independent elements fundamentally derives from two possible sources of correlation: (i) interactions between elements or groups of elements, and (ii) external influences on the system. In all human activities considered here, we witness surges of collective activity that cannot be explained under assumptions of human independence. Instead, we find that the populations are described quantitatively by models that include the simplest possible correlations – those between pairs of individuals. However, given that large-scale patterns could derive from higher-order correlations or from shared external inputs to the population, and given the myriad experiences that shape human actions, it would be naïve to universally conclude that all collective human activity emerges from pairwise correlations. Instead, we hypothesize that particular activities fall along a spectrum, with internal correlations and external influences each playing roles of variable importance.

We remark that we have already witnessed evidence for such a spectrum in the different human activities studied above. For example, while patterns of email communication were reasonably well-described by taking into account people's weekly rhythms, capturing $\sim 67\%$ of the correlation structure in 10-person groups, private message correspondence had a markedly weak dependence on people's schedules, with daily routines accounting for only $\sim 5\%$ of the correlation in 10-person groups. These results agree with intuition, indicating that email activity is moderately tied to people's work and leisure schedules, while daily routines have nearly no predictive power in a network of private messages. Interestingly, correlations in both face-to-face contacts and online music streaming are moderately driven by daily and weekly routines, falling in between email and private message correspondence. With these results in mind, the clearest direction for future investigation is to continue probing different ends of the spectrum by quantifying the relative importance of internal correlations versus external influences in different modes of human behavior.

1.8.7.2 *The energy landscape of collective human behavior*

Every maximum entropy model Q is defined by a Boltzmann distribution $Q(\boldsymbol{\sigma}) = \exp(-E(\boldsymbol{\sigma}))/Z$, where $E(\boldsymbol{\sigma})$ is the energy function, or Hamiltonian, that describes the system, and Z is the normalization constant. In the case of the pairwise maximum entropy model, the relevant energy function is that of the Ising model, $E(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{i\neq j} J_{ij}\sigma_i\sigma_j - \sum_i h_i\sigma_i$. In statistical mechanics, there is a wealth of literature exploring the diversity of large-scale behaviors that can emerge from systems with different energy landscapes (*135*, *339*). Thus, future research should leverage this connection to answer a number of important questions: What can the energy landscape of a

given population tell us about its functional properties? Does collective human behavior favor dramatic shifts in activity, or are social populations organized to incentivize local fluctuations, guarding against the effects of large external shocks?

1.8.7.3 *Beyond equal-time correlations*

Throughout our analysis, we have focused on modeling equal-time correlations, which quantify the tendencies of individuals to engage in synchronous actions. In doing so, we have implicitly assumed that each observed activity pattern $\boldsymbol{\sigma}$ is drawn independently from an underlying stationary distribution $P(\boldsymbol{\sigma})$, leaving models of the population's activity without notions of time or causality. While studying equal-time correlations has allowed us to reach a number of important conclusions, the idea that patterns of human activity are sampled from a stationary distribution is not consistent with the common intuition that conscious human actions are often responses to prior social and environmental influences. For example, the fact that individuals perform bursts of actions in quick succession is thought to be the result of a decision-based queuing process (*49*), and it is known that the temporal scales of human activity can change over time (*97*, *448*, *469*).

In the context of human communication, a significant fraction of emails and private messages are direct responses to previous correspondence. Therefore, it would be interesting to study the correlations between people's activities with a time delay τ in between, where τ represents the characteristic response time of communication in the population. Such spatiotemporal correlations have recently received a large amount of interest in neuroscience and biology, where it has been found that the spatiotemporal patterns of spiking neurons in the brain and flocks of birds in flight are only partially captured by stationary maximum entropy models (*130*, *428*, *651*). Similarly, studying the spatiotemporal patterns that define collective human activity has significant implications for understanding the causal flow of influences and information between individuals in a population (*469*). Furthermore, developing accurate dynamical models of large-scale behavior has important ramifications for predicting the effects of interventions and time-varying perturbations in networks of interacting humans (*120*, *122*, *163*, *271*, *490*, *513*, *514*).

2

MAXIMIZING ACTIVITY IN AN ISING SYSTEM: A MEAN-FIELD OPTIMAL SOLUTION

This chapter contains work from Lynn, Christopher, and Daniel D. Lee. "Maximizing influence in an Ising network: A mean-field optimal solution." Advances in Neural Information Processing Systems. 2016.

Abstract

Influence maximization in social networks has typically been studied in the context of contagion models and irreversible processes. In this paper, we consider an alternate model that treats individual opinions as spins in an Ising system at dynamic equilibrium. We formalize the *Ising influence maximization* problem, which has a natural physical interpretation as maximizing the magnetization given a budget of external magnetic field. Under the mean-field (MF) approximation, we present a gradient ascent algorithm that uses the susceptibility to efficiently calculate local maxima of the magnetization, and we develop a number of sufficient conditions for when the MF magnetization is concave and our algorithm converges to a global optimum. We apply our algorithm on random and real-world networks, demonstrating, remarkably, that the MF optimal external fields (i.e., the external fields which maximize the MF magnetization) shift from focusing on high-degree individuals at high temperatures to focusing on low-degree individuals at low temperatures. We also establish a number of novel results about the structure of steady-states in the ferromagnetic MF Ising model on general graph topologies, which are of independent interest.

2.1 INTRODUCTION

With the proliferation of online social networks, the problem of optimally influencing the opinions of individuals in a population has garnered tremendous attention (*190*, *365*, *556*). The prevailing paradigm treats marketing as a viral process, whereby the advertiser is given a budget of seed infections and chooses the subset of individuals to infect such that the spread of the ensuing contagion is maximized. The development of algorithmic methods for influence maximization under the viral paradigm has been the subject of vigorous study, resulting in a number of efficient techniques for identifying meaningful marketing strategies in real-world settings (*265*, *275*, *459*). While the viral paradigm accurately describes out-of-equilibrium phenomena, such as the introduction of new ideas or products to a system, these models fail to capture reverberant opinion

dynamics wherein repeated interactions between individuals in the network give rise to complex macroscopic opinion patterns, as, for example, is the case in the formation of political opinions (*238*, *332*, *430*, *462*). In this context, rather than maximizing the spread of a viral advertisement, the marketer is interested in optimally shifting the equilibrium opinions of individuals in the network.

To describe complex macroscopic opinion patterns resulting from repeated microscopic interactions, we naturally employ the language of statistical mechanics, treating individual opinions as spins in an Ising system at dynamic equilibrium and modeling marketing as the addition of an external magnetic field. The resulting problem, which we call *Ising influence maximization (IIM)*, has a natural physical interpretation as maximizing the magnetization of an Ising system given a budget of external field. While a number of models have been proposed for describing reverberant opinion dynamics (*173*), our use of the Ising model follows a vibrant interdisciplinary literature (*126*, *454*), and is closely related to models in game theory (*92*, *437*) and sociophysics (*237*, *648*). Furthermore, complex Ising models have found widespread use in machine learning, and our model is formally equivalent to a pair-wise Markov random field or a Boltzmann machine (*368*, *484*, *650*).

Our main contributions are as follows:

1. We formalize the influence maximization problem in the context of the Ising model, which we call the *Ising influence maximization (IIM)* problem. We also propose the *mean-field Ising influence maximization (MF-IIM)* problem as an approximation to IIM (Section 2).

2. We find sufficient conditions under which the MF-IIM objective is smooth and concave, and we present a gradient ascent algorithm that guarantees an ϵ-approximation to MF-IIM (Section 4).

3. We present numerical simulations that probe the structure and performance of MF optimal marketing strategies. We find that at high temperatures, it is optimal to focus influence on high-degree individuals, while at low temperatures, it is optimal to spread influence among low-degree individuals (Sections 5 and 6).

4. Throughout the paper we present a number of novel results concerning the structure of steady-states in the ferromagnetic MF Ising model on general (weighted, directed) strongly-connected graphs, which are of independent interest. We name two highlights:
    - The well-known pitchfork bifurcation structure for the ferromagnetic MF Ising model on a lattice extends exactly to general strongly-connected graphs, and the critical temperature is equal to the spectral radius of the adjacency matrix (Theorem 3).
    - There can exist at most one stable steady-state with non-negative (non-positive) components, and it is smooth and concave (convex) in the external field (Theorem 4).

2.2 THE ISING INFLUENCE MAXIMIZATION PROBLEM

We consider a weighted, directed social network consisting of a set of individuals $N = \{1, \dots, n\}$, each of which is assigned an opinion $\sigma_i \in \{\pm 1\}$ that captures its current state. By analogy with the Ising model, we refer to $\boldsymbol{\sigma} = (\sigma_i)$ as a spin configuration of the system. Individuals in the network interact via a non-negative weighted coupling matrix $J \in \mathbb{R}^{n \times n}_{\geqslant 0}$, where $J_{ij} \geqslant 0$ represents the amount of influence that individual j holds over the opinion of individual i, and the non-negativity of J represents the assumption that opinions of neighboring individuals tend to align, known in physics as a ferromagnetic interaction. Each individual also interacts with forces external to the network via an external field $\mathbf{h} \in \mathbb{R}^n$. For example, if the spins represent the political opinions of individuals in a social network, then J_{ij} represents the influence that j holds over i's opinion and h_i represents the political bias of node i due to external forces such as campaign advertisements and news articles.

The opinions of individuals in the network evolve according to asynchronous Glauber dynamics. At each time t, an individual i is selected uniformly at random and her opinion is updated in response to the external field $\mathbf{h}$ and the opinions of others in the network $\boldsymbol{\sigma}(t)$ by sampling from

$$P\left(\sigma_i(t+1) = 1 | \boldsymbol{\sigma}(t)\right) = \frac{e^{\beta\left(\sum_j J_{ij}\sigma_j(t) + h_i\right)}}{\sum_{\sigma_i' = \pm 1} e^{\beta\sigma_i'\left(\sum_j J_{ij}\sigma_j(t) + h_i\right)}}, \tag{2.1}$$

where β is the inverse temperature, which we refer to as the *interaction strength*, and unless otherwise specified, sums are assumed over N. Together, the quadruple $(N, J, \mathbf{h}, \beta)$ defines our system. We refer to the total expected opinion, $M = \sum_i \langle \sigma_i \rangle$, as the *magnetization*, where $\langle \cdot \rangle$ denotes an average over the dynamics in Eq. (9.5), and we often consider the magnetization as a function of the external field, denoted $M(\mathbf{h})$. Another important concept is the *susceptibility* matrix, $\chi_{ij} = \frac{\partial \langle \sigma_i \rangle}{\partial h_j}$, which quantifies the response of individual i to a change in the external field on node j.

We study the problem of maximizing the magnetization of an Ising system with respect to the external field. We assume that an external field $\mathbf{h}$ can be added to the system, subject to the constraints $\mathbf{h} \geqslant 0$ and $\sum_i h_i \leqslant H$, where $H > 0$ is the *external field budget*, and we denote the set of feasible external fields by $\mathcal{F}_H = \{\mathbf{h} \in \mathbb{R}^n : \mathbf{h} \geqslant 0, \sum_i h_i = H\}$. In general, we also assume that the system experiences an initial external field $\mathbf{b} \in \mathbb{R}^n$, which cannot be controlled.

Definition 1. (*Ising influence maximization (IIM)*) Given a system $(N, J, \mathbf{b}, \beta)$ and a budget H, find a feasible external field $\mathbf{h} \in \mathcal{F}_H$ that maximizes the magnetization; that is, find an optimal external field $\mathbf{h}^*$ such that

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathcal{F}_H} M(\mathbf{b} + \mathbf{h}). \tag{2.2}$$

NOTATION. Unless otherwise specified, bold symbols represent column vectors with the appropriate number of components, while non-bold symbols with subscripts represent individual components. We often abuse notation and write relations such as $\mathbf{m} \geqslant 0$ to mean $m_i \geqslant 0$ for all components i.

The mean-field approximation

In general, calculating expectations over the dynamics in Eq. (9.5) requires Monte-Carlo simulations or other numerical approximation techniques. To make analytic progress, we employ the variational mean-field approximation, which has roots in statistical physics and has long been used to tackle inference problems in Boltzmann machines and Markov random fields (*346*, *492*, *580*, *717*). The mean-field approximation replaces the intractable task of calculating exact averages over Eq. (9.5) with the problem of solving the following set of self-consistency equations:

$$m_i = \tanh\left[\beta\left(\sum_j J_{ij} m_j + h_i\right)\right], \tag{2.3}$$

for all $i \in N$, where m_i approximates $\langle \sigma_i \rangle$. We refer to the right-hand side of Eq. (2.3) as the *mean-field map*, $\mathbf{f}(\mathbf{m}) = \tanh[\beta(J\mathbf{m} + \mathbf{h})]$, where $\tanh(\cdot)$ is applied component-wise. In this way, a fixed point of the mean-field map is a solution to Eq. (2.3), which we call a *steady-state*.

In general, there may be many solutions to Eq. (2.3), and we denote by $\mathcal{M}_{\mathbf{h}}$ the set of steady-states for a system $(N, J, \mathbf{h}, \beta)$. We say that a steady-state $\mathbf{m}$ is *stable* if $\rho(\mathbf{f}'(\mathbf{m})) < 1$, where $\rho(\cdot)$ denotes the spectral radius and

$$\mathbf{f}'(\mathbf{m})_{ij} = \left.\frac{\partial f_i}{\partial m_j}\right|_{\mathbf{m}} = \beta\left(1 - m_i^2\right) J_{ij} \quad \Rightarrow \quad \mathbf{f}'(\mathbf{m}) = \beta D(\mathbf{m}) J, \tag{2.4}$$

where $D(\mathbf{m})_{ij} = (1 - m_i^2)\delta_{ij}$. Furthermore, under the mean-field approximation, given a stable steady-state $\mathbf{m}$, the susceptibility has a particularly nice form:

$$\chi_{ij}^{MF} = \beta\left(1 - m_i^2\right)\left(\sum_k J_{ik}\chi_{kj} + \delta_{ij}\right) \quad \Rightarrow \quad \chi^{MF} = \beta\left(I - \beta D(\mathbf{m}) J\right)^{-1} D(\mathbf{m}), \tag{2.5}$$

where I is the $n \times n$ identity matrix.

For the purpose of uniquely defining our objective, we optimistically choose to maximize the maximum magnetization among the set of steady-states, defined by

$$M^{MF}(\mathbf{h}) = \max_{\mathbf{m} \in \mathcal{M}_{\mathbf{h}}} \sum_i m_i(\mathbf{h}). \tag{2.6}$$

We note that the pessimistic framework of maximizing the minimum magnetization yields an equally valid objective. We also note that simply choosing a steady-state to

optimize does not yield a well-defined objective since, as $\mathbf{h}$ increases, steady-states can pop in and out of existence.

Definition 2. (*Mean-field Ising influence maximization (MF-IIM)*) Given a system $(N, J, \mathbf{b}, \beta)$ and a budget H, find an optimal external field $\mathbf{h}^*$ such that

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathcal{F}_H} M^{MF}(\mathbf{b} + \mathbf{h}). \tag{2.7}$$

2.3 THE STRUCTURE OF STEADY-STATES IN THE MF ISING MODEL

Before proceeding further, we must prove an important result concerning the existence and structure of solutions to Eq. (2.3), for if there exists a system that does not admit a steady-state, then our objective is ill-defined. Furthermore, if there exists a unique steady-state $\mathbf{m}$, then $M^{MF} = \sum_i m_i$, and there is no ambiguity in our choice of objective.

Theorem 3 establishes that every system admits a steady-state and that the well-known pitchfork bifurcation structure for steady-states of the ferromagnetic MF Ising model on a lattice extends exactly to general (weighted, directed) strongly-connected graphs. In particular, for any strongly-connected graph described by J, there is a *critical interaction strength* β_c below which there exists a unique and stable steady-state. For $\mathbf{h} = 0$, as β crosses β_c from below, two new stable steady-states appear, one with all-positive components and one with all-negative components. Interestingly, the critical interaction strength is equal to the inverse of the spectral radius of J, denoted $\beta_c = 1/\rho(J)$.

Theorem 3. *Any system $(N, J, \mathbf{h}, \beta)$ exhibits a steady-state. Furthermore, if its network is strongly-connected, then, for $\beta < \beta_c$, there exists a unique and stable steady-state. For $\mathbf{h} = 0$, as β crosses β_c from below, the unique steady-state gives rise to two stable steady-states, one with all-positive components and one with all-negative components.*

Proof sketch. The existence of a steady-state follows directly by applying Brouwer's fixed-point theorem to $\mathbf{f}$. For $\beta < \beta_c$, it can be shown that $\mathbf{f}$ is a contraction mapping, and hence admits a unique and stable steady-state by Banach's fixed point theorem. For $\mathbf{h} = 0$ and $\beta < \beta_c$, $\mathbf{m} = 0$ is the unique steady-state and $\mathbf{f}'(\mathbf{m}) = \beta J$. Because J is strongly-connected, the Perron-Frobenius theorem guarantees a simple eigenvalue equal to $\rho(J)$ and a corresponding all-positive eigenvector. Thus, when β crosses $1/\rho(J)$ from below, the Perron-Frobenius eigenvalue of $\mathbf{f}'(\mathbf{m})$ crosses 1 from below, giving rise to a supercritical pitchfork bifurcation with two new stable steady-states corresponding to the Perron-Frobenius eigenvector.

Remark. Some of our results assume J is strongly-connected in order to use the Perron-Frobenius theorem. We note that this assumption is not restrictive, since any graph can

be efficiently decomposed into strongly-connected components on which our results apply independently.

Theorem 3 shows that the objective $M^{MF}(\mathbf{b}+\mathbf{h})$ is well-defined. Furthermore, for $\beta < \beta_c$, Theorem 3 guarantees a unique and stable steady-state $\mathbf{m}$ for all $\mathbf{b}+\mathbf{h}$. In this case, MF-IIM reduces to maximizing $M^{MF} = \sum_i m_i$, and because $\mathbf{m}$ is stable, $M^{MF}(\mathbf{b}+\mathbf{h})$ is smooth for all $\mathbf{h}$ by the implicit function theorem. Thus, for $\beta < \beta_c$, we can use standard gradient ascent techniques to efficiently calculate locally-optimal solutions to MF-IIM. In general, M^{MF} is not necessarily smooth in $\mathbf{h}$ since the topological structure of steady-states may change as $\mathbf{h}$ varies. However, in the next section we show that if there exists a stable and entry-wise non-negative steady-state, and if J is strongly-connected, then $M^{MF}(\mathbf{b}+\mathbf{h})$ is both smooth and concave in $\mathbf{h}$, regardless of the interaction strength.

2.4 SUFFICIENT CONDITIONS FOR WHEN MF-IIM IS CONCAVE

We consider conditions for which MF-IIM is smooth and concave, and hence exactly solvable by efficient techniques. The case under consideration is when J is strongly-connected and there exists a stable non-negative steady-state.

Theorem 4. *Let* $(N, J, \mathbf{b}, \beta)$ *describe a system with a strongly-connected graph for which there exists a stable non-negative steady-state* $\mathbf{m}(\mathbf{b})$*. Then, for any* H, $M^{MF}(\mathbf{b}+\mathbf{h}) = \sum_i m_i(\mathbf{b}+\mathbf{h})$, $M^{MF}(\mathbf{b}+\mathbf{h})$ *is smooth in* $\mathbf{h}$*, and* $M^{MF}(\mathbf{b}+\mathbf{h})$ *is concave in* $\mathbf{h}$ *for all* $\mathbf{h} \in \mathcal{F}_H$.

Proof sketch. Our argument follows in three steps. We first show that $\mathbf{m}(\mathbf{b})$ is the unique stable non-negative steady-state and that it attains the maximum total opinion among steady-states. This guarantees that $M^{MF}(\mathbf{b}) = \sum_i m_i(\mathbf{b})$. Furthermore, $\mathbf{m}(\mathbf{b})$ gives rise to a unique and smooth branch of stable non-negative steady-states for additional $\mathbf{h}$, and hence $M^{MF}(\mathbf{b}+\mathbf{h}) = \sum_i m_i(\mathbf{b}+\mathbf{h})$ for all $\mathbf{h} > 0$. Finally, one can directly show that $M^{MF}(\mathbf{b}+\mathbf{h})$ is concave in $\mathbf{h}$.

Remark. By arguments similar to those in Theorem 4, it can be shown that any stable non-positive steady-state is unique, attains the minimum total opinion among steady-states, and is smooth and convex for decreasing $\mathbf{h}$.

The above result paints a significantly simplified picture of the MF-IIM problem when J is strongly-connected and there exists a stable non-negative steady-state $\mathbf{m}(\mathbf{b})$. Given a budget H, for any feasible marketing strategy $\mathbf{h} \in \mathcal{F}_H$, $\mathbf{m}(\mathbf{b}+\mathbf{h})$ is the unique stable non-negative steady-state, attains the maximum total opinion among steady-states, and is smooth in $\mathbf{h}$. Thus, the objective $M^{MF}(\mathbf{b}+\mathbf{h}) = \sum_i m_i(\mathbf{b}+\mathbf{h})$ is smooth, allowing us to write down a gradient ascent algorithm that approximates a local maximum. Furthermore, since $M^{MF}(\mathbf{b}+\mathbf{h})$ is concave in $\mathbf{h}$, any local maximum of M^{MF} on $\mathcal{F}_H$

Algorithm 1: An ϵ-approximation to MF-IIM

Input: System $(N, J, \mathbf{b}, \beta)$ for which there exists a stable non-negative steady-state, budget H, accuracy parameter $\epsilon > 0$
Output: External field $\mathbf{h}$ that approximates a MF optimal external field $\mathbf{h}^*$
$t = 0$; $\mathbf{h}(0) \in \mathcal{F}_H$; $\alpha \in (0, \frac{1}{L})$;
repeat
  $\frac{\partial M^{MF}(\mathbf{b}+\mathbf{h}(t))}{\partial h_j} = \sum_i \chi_{ij}^{MF}(\mathbf{b} + \mathbf{h}(t))$;
  $\mathbf{h}(t+1) = P_{\mathcal{F}_H}\left[\mathbf{h}(t) + \alpha \nabla_{\mathbf{h}} M^{MF}(\mathbf{b} + \mathbf{h}(t))\right]$;
  $t{+}{+}$;
until $M^{MF}(\mathbf{b} + \mathbf{h}^*) - M^{MF}(\mathbf{b} + \mathbf{h}(t)) \leqslant \epsilon$;
$\mathbf{h} = \mathbf{h}(t)$;

is a global maximum, and we can apply efficient gradient ascent techniques to solve MF-IIM.

Our algorithm, summarized in Algorithm 1, is initialized at a feasible external field. At each iteration, we calculate the susceptibility of the system, namely $\frac{\partial M^{MF}}{\partial h_j} = \sum_i \chi_{ij}^{MF}$, and project this gradient onto $\mathcal{F}_H$ (the projection operator $P_{\mathcal{F}_H}$ is well-defined since $\mathcal{F}_H$ is convex). Stepping along the direction of the projected gradient with step size $\alpha \in (0, \frac{1}{L})$, where L is a Lipschitz constant of M^{MF}, Algorithm 1 converges to an ϵ-approximation to MF-IIM in $O(1/\epsilon)$ iterations (*657*).

Sufficient conditions for the existence of a stable non-negative steady-state

In the previous section we found that MF-IIM is efficiently solvable if there exists a stable non-negative steady-state. While this assumption may seem restrictive, we show, to the contrary, that the appearance of a stable non-negative steady-state is a fairly general phenomenon. We first show, for J strongly-connected, that the existence of a stable non-negative steady-state is robust to increases in $\mathbf{h}$ and that the existence of a stable positive steady-state is robust to increases in β.

Theorem 5. *Let $(N, J, \mathbf{h}, \beta)$ describe a system with a strongly-connected graph for which there exists a stable non-negative steady-state $\mathbf{m}$. If $\mathbf{m} \geqslant 0$, then as $\mathbf{h}$ increases, $\mathbf{m}$ gives rise to a unique and smooth branch of stable non-negative steady-states. If $\mathbf{m} > 0$, then as β increases, $\mathbf{m}$ gives rise to a unique and smooth branch of stable positive steady-states.*

Proof sketch. By the implicit function theorem, any stable steady-state can be locally defined as a function of both $\mathbf{h}$ and β. Using the susceptibility, one can directly show that any stable non-negative steady-state remains stable and non-negative as $\mathbf{h}$ increases and that any stable positive steady-state remains stable and positive as β increases.

The intuition behind Theorem 5 is that increasing the external field will never destroy a steady-state in which all of the opinions are already non-positive. Furthermore, as

the interaction strength increases, each individual reacts more strongly to the positive influence of her neighbors, creating a positive feedback loop that results in an even more positive magnetization. We conclude by showing for J strongly-connected that if $\mathbf{h} \geqslant 0$, then there exists a stable non-negative steady-state.

Theorem 6. *Let* $(N, J, \mathbf{h}, \beta)$ *describe any system with a strongly-connected network. If* $\mathbf{h} \geqslant 0$, *then there exists a stable non-negative steady-state.*

Proof sketch. For $\mathbf{h} > 0$ and $\beta < \beta_c$, it can be shown that the unique steady-state is positive, and hence Theorem 5 guarantees the result for all $\beta' > \beta$. For $\mathbf{h} = 0$, Theorem 3 provides the result.

All together, the results of this section provide a number of sufficient conditions under which MF-IIM is exactly and efficiently solvable by Algorithm 1.

2.5 A SHIFT IN THE STRUCTURE OF SOLUTIONS TO MF-IIM

The structure of solutions to MF-IIM is of fundamental theoretical and practical interest. We demonstrate, remarkably, that solutions to MF-IIM shift from focusing on nodes of high degree at low interaction strengths to focusing on nodes of low degree at high interaction strengths.

Consider an Ising system described by $(N, J, \mathbf{h}, \beta)$ in the limit $\beta \ll \beta_c$. To first-order in β, the self-consistency equations (2.3) take the form:

$$\mathbf{m} = \beta\,(J\mathbf{m} + \mathbf{h}) \quad \Rightarrow \quad \mathbf{m} = \beta(I - \beta J)^{-1}\mathbf{h}. \tag{2.8}$$

Since $\beta < \beta_c$, we have $\rho(\beta J) < 1$, allowing us to expand $(I - \beta J)^{-1}$ in a geometric series:

$$\mathbf{m} = \beta\mathbf{h} + \beta^2 J\mathbf{h} + O(\beta^3) \quad \Rightarrow \quad M^{MF}(\mathbf{h}) = \beta \sum_i h_i + \beta^2 \sum_i d_i^{out} h_i + O(\beta^3), \tag{2.9}$$

where $d_i^{out} = \sum_j J_{ji}$ is the out-degree of node i. Thus, for low interaction strengths, the MF magnetization is maximized by focusing the external field on the nodes of highest out-degree in the network, independent of $\mathbf{b}$ and H.

To study the structure of solutions to MF-IIM at high interaction strengths, we make the simplifying assumptions that J is strongly-connected and $\mathbf{b} \geqslant 0$ so that Theorem 6 guarantees a stable non-negative steady state $\mathbf{m}$. For large β and an additional external field $\mathbf{h} \in \mathcal{F}_H$, $\mathbf{m}$ takes the form

$$m_i \approx \tanh\left[\beta\left(\sum_j J_{ij} + b_i + h_i\right)\right] \approx 1 - 2e^{-2\beta(d_i^{in} + b_i + h_i)}, \tag{2.10}$$

where $d_i^{in} = \sum_j J_{ij}$ is the in-degree of node i. Thus, in the high-β limit, we have:

$$M^{MF}(\mathbf{b}+\mathbf{h}) \approx \sum_i \left(1 - 2e^{-2\beta(d_i^{in}+b_i+h_i)}\right) \approx n - 2e^{-2\beta(d_{i^*}^{in}+h_{i^*}^{(0)}+h_{i^*})}, \tag{2.11}$$

where $i^* = \arg\min_i(d_i^{in} + b_i + h_i)$. Thus, for high interaction strengths, the solutions to MF-IIM for an external field budget H are given by:

$$\mathbf{h}^* = \arg\max_{\mathbf{h}\in\mathcal{F}_H} \left(n - 2e^{-2\beta(d_{i^*}^{in}+h_{i^*}^{(0)}+h_{i^*})}\right) \equiv \arg\max_{\mathbf{h}\in\mathcal{F}_H} \min_i \left(d_i^{in} + b_i + h_i\right). \tag{2.12}$$

Eq. (2.12) reveals that the high-β solutions to MF-IIM focus on the nodes for which $d_i^{in} + b_i + h_i$ is smallest. Thus, if $\mathbf{b}$ is uniform, the MF magnetization is maximized by focusing the external field on the nodes of smallest in-degree in the network.

We emphasize the strength and novelty of the above results. In the context of reverberant opinion dynamics, the optimal control strategy has a highly non-trivial dependence on the strength of interactions in the system, a feature not captured by viral models. Thus, when controlling a social system, accurately determining the strength of interactions is of critical importance.

2.6 NUMERICAL SIMULATIONS

We present numerical experiments to probe the structure and performance of MF optimal external fields. We verify that the solutions to MF-IIM undergo a shift from focusing on high-degree nodes at low interaction strengths to focusing on low-degree nodes at high interaction strengths. We also find that for sufficiently high and low interaction strengths, the MF optimal external field achieves the maximum exact magnetization, while admitting performance losses near β_c. However, even at β_c, we demonstrate that solutions to MF-IIM significantly outperform common node-selection heuristics based on node degree and centrality.

We first consider an undirected hub-and-spoke network, shown in Fig. 2.1a, where $J_{ij} \in \{0, 1\}$ and we set $\mathbf{b} = 0$ for simplicity. Since $\mathbf{b} \geqslant 0$, Algorithm 1 is guaranteed to achieve a globally optimal MF magnetization. Furthermore, because the network is small, we can calculate exact solutions to IIM by brute force search. In Fig. 2.1a, we compare the average degree of the MF and exact optimal external fields over a range of temperatures for an external field budget $H = 1$, verifying that the solution to MF-IIM shifts from focusing on high-degree nodes at low interaction strengths to low-degree nodes at high interaction strengths. Furthermore, we find that the shift in the MF optimal external field occurs near the critical interaction strength $\beta_c = .5$. The performance of the MF optimal strategy (measured as the ratio of the magnetization achieved by the MF solution to that achieved by the exact solution) is shown in Fig. 2.1b. For low and high interaction strengths, the MF optimal external field achieves the maximum magnetization, while near β_c, it incurs significant performance losses, a phenomenon well-studied in the literature (*717*).

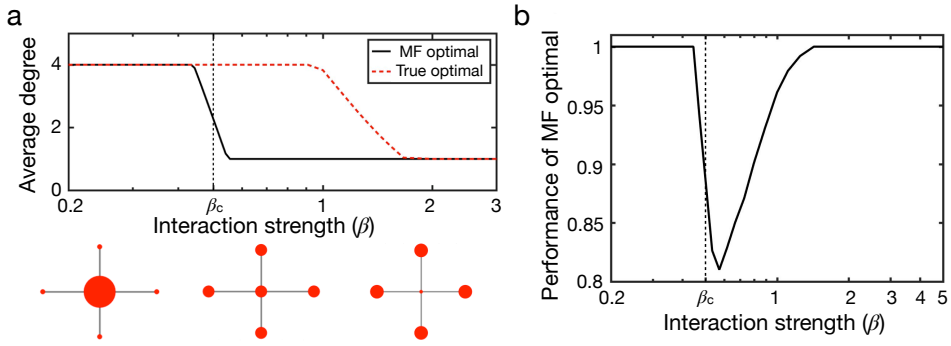


Figure 2.1: Optimal and MF optimal external fields for a hub-and-spoke network. (*a*) A comparison of the structure of the MF and exact optimal external fields, denoted $\mathbf{h}^*_{\mathrm{MF}}$ and $\mathbf{h}^*$, in a hub-and-spoke network. (**b**) The relative performance of $\mathbf{h}^*_{\mathrm{MF}}$ compared to $\mathbf{h}^*$; i.e., $M(\mathbf{h}^*_{\mathrm{MF}})/M(\mathbf{h}^*_{\mathrm{MF}})$, where M denotes the exact magnetization.

We now consider a stochastic block network consisting of 100 nodes split into two blocks of 50 nodes each, shown in Fig. 2.2a. An undirected edge of weight 1 is placed between each pair of nodes in Block 1 with probability .2, between each pair in Block 2 with probability .05, and between nodes in different blocks with probability .05, resulting in a highly-connected community (Block 1) surrounded by a sparsely-connected community (Block 2). For $\mathbf{b} = 0$ and $H = 20$, Fig. 2.2b demonstrates that the solution to MF-IIM shifts from focusing on Block 1 at low β to focusing on Block 2 at high β and that the shift occurs near β_c. The stochastic block network is sufficiently large that exact calculation of the optimal external fields is infeasible. Thus, we resort to comparing the MF solutions with three node-selection heuristics: one that distributes the budget in amounts proportional to nodes' degrees, one that distributes the budget proportional to nodes' centralities (the inverse of a node's average shortest path length to all other nodes), and one that distributes the budget randomly. The magnetizations are approximated via Monte Carlo simulations of the Glauber dynamics, and we consider the system at $\beta = \beta_c$ to represent the worst-case scenario for the MF optimal external fields. In Fig. 2.2c, we show that, even at β_c, the solutions to MF-IIM outperform common node-selection heuristics.

We consider a real-world collaboration network (Fig. 2.3a) composed of 904 individuals, where each edge is unweighted and represents the co-authorship of a paper on the arXiv (*395*). We note that co-authorship networks are known to capture many of the key structural features of social networks (*477*). For $\mathbf{b} = 0$ and $H = 40$, Fig. 2.3b illustrates the sharp shift in the solution to MF-IIM at $\beta_c = 0.05$ from high- to low-degree nodes. Furthermore, in Fig. 2.3c we compare the performance of the MF optimal external field with the node-selection heuristics described above, where we again consider the system at β_c as a worst-case scenario, demonstrating that Algorithm 1 is scalable and performs well on real-world networks.

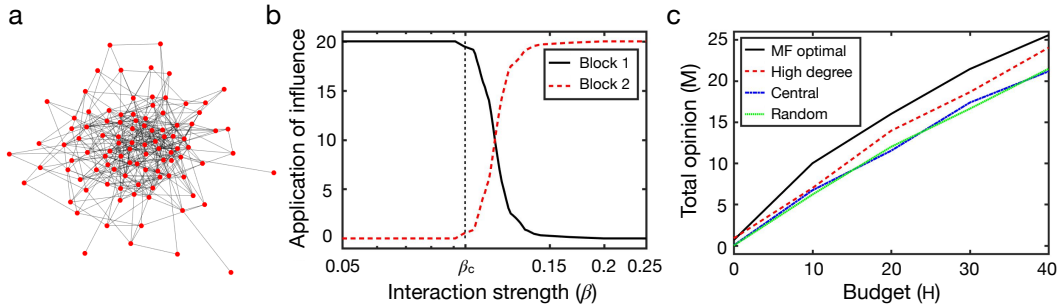


Figure 2.2: Structure of MF optimal external field for a stochastic block network. (*a*) A stochastic block network consisting of a highly-connected community (Block 1) and a sparsely-connected community (Block 2). (*b*) The solution to MF-IIM shifts from focusing on Block 1 to Block 2 as β increases. (*c*) Even at β_c, the MF solution outperforms common node-selection heuristics.

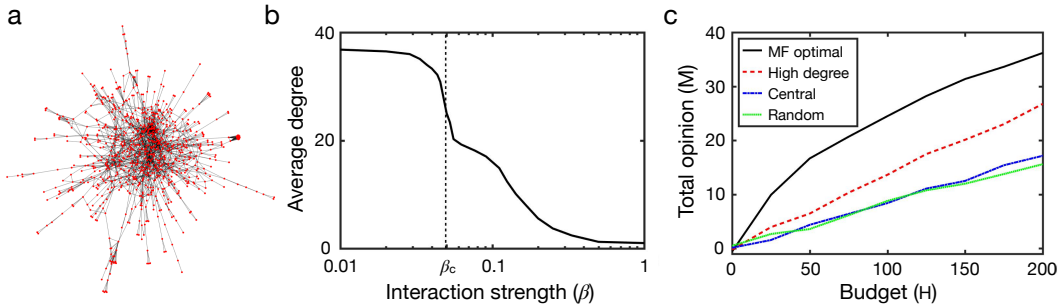


Figure 2.3: Structure of MF optimal external field for real-world social network. (*a*) A collaboration network of 904 physicists where each edge represents the co-authorship of a paper on the arXiv. (*b*) The solution to MF-IIM shifts from high- to low-degree nodes as β increases. (*c*) The MF solution out-performs common node-selection heuristics, even at β_c.

2.7 CONCLUSIONS

We study influence maximization, one of the fundamental problems in network science, in the context of the Ising model, wherein repeated interactions between individuals give rise to complex macroscopic patterns. The resulting problem, which we call Ising influence maximization, has a natural physical interpretation as maximizing the magnetization of an Ising system given a budget of external magnetic field. Under the mean-field approximation, we develop a number of sufficient conditions for when the problem is concave, and we provide a gradient ascent algorithm that uses the susceptibility to efficiently calculate locally-optimal external fields. Furthermore, we demonstrate that the MF optimal external fields shift from focusing on high-degree individuals at low interaction strengths to focusing on low-degree individuals at high interaction strengths, a phenomenon not observed in viral models. We apply our algorithm on random and real-world networks, numerically demonstrating shifts in the solution structure and showing that our algorithm out-performs common node-selection heuristics.

It would be interesting to study the exact Ising model on an undirected network, in which case the spin statistics are governed by the Boltzmann distribution. Using this elegant steady-state description, one might be able to derive analytic results for the exact IIM problem. Our work establishes a fruitful connection between influence maximization and statistical physics, paving the way for exciting cross-disciplinary research. For example, one could apply advanced mean-field techniques, such as those in (*717*), to generate efficient algorithms of increasing accuracy. Furthermore, because our model is equivalent to a Boltzmann machine, one could propose a framework for data-based influence maximization based on well-known Boltzmann machine learning techniques.

2.8 SUPPLEMENTARY MATERIAL

2.8.1 *Preliminaries*

We establish a number of preliminary results that aid the proofs of the theorems in the main text.

2.8.1.1 *Perron-Frobenius*

Many of the results in the paper rely on the Perron-Frobenius theorem, which we state here.

Theorem 7. (Perron-Frobenius) *Let J be an irreducible non-negative matrix with spectral radius $\rho(J) = r$. Then the following statements hold:*

1. *J has a real, positive, and simple eigenvalue equal to r.*
2. *The corresponding eigenvector of J has all-positive components.*
3. *If $0 \leqslant J \leqslant A$, for some matrix A, then $r_J \leqslant r_A$.*

It is known that the adjacency matrix of a strongly-connected graph is irreducible, and hence, all of the results of the Perron-Frobenius theorem carry over.

2.8.1.2 *The existence of a unique and stable steady-state for $\beta < \beta_c$*

We first show that if our network is strongly-connected, then for $\beta < 1/\rho(J)$, the system exhibits a unique steady-state that is stable under $\mathbf{f}$. This result will aid in the proof of Theorem 3 and similar arguments are used in the proof of Theorem 4. The proof in this section is based on Banach's fixed point theorem, but instead of directly showing that $\mathbf{f}$ is a contraction mapping on $X = [-1, 1]^n$, we use the spectral properties of $\mathbf{f}'$. The following two lemmas relate the contraction mapping property to the spectral radius of $\mathbf{f}'$. We note that throughout the proofs, we use the variable $\mathbf{x}$ instead of $\mathbf{m}$ to indicate a point in X that is not necessarily a steady-state of $\mathbf{f}$.

Lemma 8. *Let X be a convex subset of Euclidean space and let the function $\mathbf{f} : X \to X$ have continuous partial derivatives on X. If the Jacobian satisfies*

$$|\mathbf{f}'(\mathbf{x})| < 1, \tag{2.13}$$

for all $\mathbf{x} \in X$ and some matrix norm $|\cdot|$, then $\mathbf{f}$ satisfies the contraction mapping property on X.

Lemma 9. *Given a square matrix A and $\epsilon > 0$, there exists a matrix norm $|\cdot|$ such that*

$$|A| \leqslant \rho(A) + \epsilon. \tag{2.14}$$

We are now prepared to show that if J is strongly-connected, then for $\beta < 1/\rho(J)$, $\mathbf{f}$ is a contraction mapping on X, and hence $\mathbf{f}$ admits a unique and stable fixed point on X.

Lemma 10. *Let* $(N, J, \mathbf{h}, \beta)$ *describe a system with a strongly-connected graph. For* $\beta < \beta_c$, *there exists a unique and stable steady-state that can be found by iteratively applying* $\mathbf{f}$ *to any point* $\mathbf{x} \in X$.

Proof. Consider the Jacobian,

$$\mathbf{f}'(\mathbf{m})_{ij} = \frac{\partial}{\partial m_j} \tanh\left[\beta(J\mathbf{m} + \mathbf{h})\right]_i = \beta \operatorname{sech}^2\left[\beta(J\mathbf{m} + \mathbf{h})\right]_i J_{ij}. \tag{2.15}$$

Since $|\operatorname{sech}(\cdot)| \leqslant 1$, $|\mathbf{f}'(\mathbf{m})_{ij}| \leqslant \beta J_{ij}$ for all $i, j \in N$. For $\beta < 1/\rho(J)$ and for all $\mathbf{m} \in X$, we have

$$\rho(\mathbf{f}'(\mathbf{m})) \leqslant \rho(\beta J) = \beta\rho(J) < 1, \tag{2.16}$$

where the first inequality follows from the Perron-Frobenius theorem and the equality follows from the linearity of $\rho(\cdot)$.

Because the above inequality is strict, there exists an $\epsilon > 0$ such that $\rho(\mathbf{f}'(\mathbf{x})) + \epsilon < 1$ for all $\mathbf{x} \in X$. By Lemma 9 we can choose a matrix norm $|\cdot|$ such that

$$|\mathbf{f}'(\mathbf{m})| \leqslant \rho(\mathbf{f}'(\mathbf{m})) + \epsilon < 1. \tag{2.17}$$

Since X is a convex subset of Euclidean space, Lemma 8 implies that $\mathbf{f}$ satisfies the contraction mapping property on X. Since X is a closed and bounded it is also a compact metric space and we can apply Banach's theorem on compact metric spaces to attain the desired result. □

2.8.1.3 *The smoothness of stable non-negative steady-states for increasing* h

We show that any stable non-negative steady-state $\mathbf{m}$ gives rise to a unique and stable branch of steady-states that is smooth and non-decreasing as $\mathbf{h}$ increases. We note that this result represents half of the progress toward proving Theorem 5.

Lemma 11. *Let* $(N, J, \mathbf{h}, \beta)$ *describe a system with a strongly-connected graph for which there exists a stable steady-state* $\mathbf{m}$. *If* $\mathbf{m} \geqslant 0$, *then as* $\mathbf{h}$ *increases,* $\mathbf{m}$ *gives rise to a unique and smooth branch of stable non-negative steady-states.*

Proof. We first show that any stable steady-state is locally non-decreasing in $\mathbf{h}$. Consider the susceptibility from Sec. 2:

$$\chi^{MF} = \beta(I - \beta D(\mathbf{m})J)^{-1} D(\mathbf{m}) = \beta(I - \mathbf{f}'(\mathbf{m}))^{-1} D(\mathbf{m}). \tag{2.18}$$

Since $\rho(\mathbf{f}'(\mathbf{m})) < 1$, Theorem 4.3 of (*216*) guarantees that the matrix $(I - \mathbf{f}'(\mathbf{m}))^{-1}$ is non-negative. Furthermore, since $D(\mathbf{m})$ is non-negative, we find that χ^{MF} is non-

negative, and hence $\frac{\partial m_i}{\partial h_j} \geqslant 0$ for all $i, j \in N$.

We now argue that $\mathbf{m}$ remains stable as $\mathbf{h}$ increases which, by the implicit function theorem, guarantees that $\mathbf{m}$ branches uniquely and smoothly. It is sufficient to show that $\rho(\mathbf{f}'(\mathbf{m})) = \rho(\beta D(\mathbf{m})J)$ is non-increasing in $\mathbf{h}$. Since $\mathbf{m}$ is non-decreasing in $\mathbf{h}$, we find that $\mathbf{f}'(\mathbf{m})$ is entry-wise non-increasing in $\mathbf{h}$, and hence Perron-Frobenius guarantees that $\rho(\mathbf{f}'(\mathbf{m}))$ is non-increasing. □

2.8.2 *Proofs*

We now present proofs of the theorems in the main text, noting that we do not present the results in the order that they appear in the text since some earlier results depend on later ones.

2.8.2.1 *Theorem 4*

We split Theorem 4 into three separate results. We first show that any stable non-negative steady-state is the unique stable non-negative steady state and that it attains the maximum total opinion among all steady-states. Secondly, we note that Lemma 11 guarantees that any stable non-negative steady-state is smooth and remains stable and non-negative for increasing $\mathbf{h}$. Finally, we show that any stable non-negative steady-state is concave in $\mathbf{h}$. Together, these results prove Theorem 4.

Uniqueness of stable non-negative steady-states

We show that any stable non-negative steady-state is the unique stable non-negative steady-state and achieves the largest total opinion among steady-states. First consider the following lemma.

Lemma 12. *Let* $(N, J, \mathbf{h}, \beta)$ *describe an arbitrary system and consider any point* $\mathbf{x} \in X$. *Then*

$$\mathbf{f}^{\ell}(\mathbf{1}) \geqslant \mathbf{f}^{\ell}(\mathbf{x}), \tag{2.19}$$

for any positive integer ℓ, *where* $\mathbf{f}^{\ell}(\cdot)$ *denotes the* ℓ^{th} *iterative application of* $\mathbf{f}$ *and* $\mathbf{1}$ *is the vector of ones of length* n.

Proof. We proceed by induction. The base case is trivially satisfied. For the inductive step, assume $f^{\ell}(\mathbf{1})_i \geqslant f^{\ell}(\mathbf{x})_i$ for some ℓ and all $i \in N$. Since $J \geqslant 0$, $\beta \geqslant 0$ and $\tanh(\cdot)$ is increasing, we have

$$f^{\ell+1}(\mathbf{1})_i = \tanh\left[\beta\left(Jf^{\ell}(\mathbf{1}) + \mathbf{h}\right)\right]_i \geqslant \tanh\left[\beta\left(Jf^{\ell}(\mathbf{x}) + \mathbf{h}\right)\right]_i = f^{\ell+1}(\mathbf{x})_i, \tag{2.20}$$

for all $\mathbf{x} \in X$ and all $i \in N$. □

We now establish the uniqueness of stable non-negative steady-states.

Lemma 13. *Let* $(N, J, \mathbf{h}, \beta)$ *describe a system with a strongly-connected network for which there exists a stable non-negative steady-state* $\mathbf{m}$*. Then* $\mathbf{m}$ *is the unique stable steady-state and can be found by iteratively applying* $\mathbf{f}$ *to* $\mathbf{1}$*.*

Proof. Assume there exists a stable non-negative steady-state $\mathbf{m}$. By Lemma 12,

$$\mathbf{f}^{\ell}(\mathbf{1}) \geqslant \mathbf{f}^{\ell}(\mathbf{m}) = \mathbf{m}, \tag{2.21}$$

for any ℓ. This indicates that the sequence $\{\mathbf{f}^{\ell}(\mathbf{1})\}$ is contained in the closed region $X_{\mathbf{m}} = \{\mathbf{x} \in X : \mathbf{x} \geqslant \mathbf{m}\}$. By an argument similar to that in the proof of Lemma 11, we have $\rho(\mathbf{f}'(\mathbf{x})) \leqslant \rho(\mathbf{f}'(\mathbf{m})) < 1$ for all $\mathbf{x} \in X_{\mathbf{m}}$. By an argument similar to that in the proof of Lemma 10, this indicates that $\mathbf{f}$ is a contraction mapping on $X_{\mathbf{m}}$.

Following the proof in (*498*), we show that the sequence $\{\mathbf{f}^{\ell}(\mathbf{1})\}$ is Cauchy. Since $|\mathbf{f}(\mathbf{x}') - \mathbf{f}(\mathbf{x})| < |\mathbf{x}' - \mathbf{x}|$ for all $\mathbf{x}, \mathbf{x}' \in X_{\mathbf{m}}$, there exists a number $q \in (0, 1)$ such that $|\mathbf{f}(\mathbf{x}') - \mathbf{f}(\mathbf{x})| \leqslant q|\mathbf{x}' - \mathbf{x}|$. By the triangle inequality,

$$|\mathbf{x}' - \mathbf{x}| \leqslant |\mathbf{x}' - \mathbf{f}(\mathbf{x}')| + q|\mathbf{x}' - \mathbf{x}| + |\mathbf{f}(\mathbf{x}) - \mathbf{x}|, \tag{2.22}$$

which yields

$$|\mathbf{x}' - \mathbf{x}| \leqslant \frac{|\mathbf{f}(\mathbf{x}') - \mathbf{x}'| + |\mathbf{f}(\mathbf{x}) - \mathbf{x}|}{1 - q}. \tag{2.23}$$

Replacing $\mathbf{x}$ and $\mathbf{x}'$ with $\mathbf{f}^{\ell}(\mathbf{1})$ and $\mathbf{f}^{k}(\mathbf{1})$, respectively, we find

$$\begin{aligned} |\mathbf{f}^{k}(\mathbf{1}) - \mathbf{f}^{\ell}(\mathbf{1})| &\leqslant \frac{|\mathbf{f}^{k+1}(\mathbf{1}) - \mathbf{f}^{k}(\mathbf{1})| + |\mathbf{f}^{\ell+1}(\mathbf{1}) - \mathbf{f}^{\ell}(\mathbf{1})|}{1 - q} \\ &\leqslant \frac{q^{k} + q^{\ell}}{1 - q} |\mathbf{f}(\mathbf{1}) - \mathbf{1}|. \end{aligned} \tag{2.24}$$

Since $q < 1$, the last expression goes to zero as $\ell, k \to \infty$, proving that $\{\mathbf{f}^{\ell}(\mathbf{1})\}$ is Cauchy and hence converges to a limit $\mathbf{m}^* \in X_{\mathbf{m}}$. Furthermore, the limit $\mathbf{m}^*$ is a fixed point of $\mathbf{f}$, and hence a steady-state of the system, since

$$\mathbf{m}^* = \lim_{\ell \to \infty} \mathbf{f}^{\ell}(\mathbf{1}) = \lim_{\ell \to \infty} \mathbf{f}(\mathbf{f}^{\ell-1}(\mathbf{1})) = \mathbf{f}\left(\lim_{\ell \to \infty} \mathbf{f}^{\ell-1}(\mathbf{1})\right) = \mathbf{f}(\mathbf{m}^*). \tag{2.25}$$

Suppose for contradiction that $\mathbf{m}^* \neq \mathbf{m}$, and consider the line $(1-t)\mathbf{m} + t\mathbf{m}^*$ between $\mathbf{m}$ and $\mathbf{m}^*$ for $t \in [0,1]$. All points along this line lie in $X_{\mathbf{m}}$ and hence $\mathbf{f}$ is contractive along the line. We have,

$$\begin{aligned} |\mathbf{f}(\mathbf{m}^*) - \mathbf{f}(\mathbf{m})| &= \left| \int_{\mathbf{m}}^{\mathbf{m}^*} \mathbf{f}'(\mathbf{x}) \cdot d\mathbf{x} \right| \\ &\leqslant \int_0^1 \left| \mathbf{f}'\left((1-t)\mathbf{m} + t\mathbf{m}^*\right) \right| |\mathbf{m}^* - \mathbf{m}| \, dt, \end{aligned} \tag{2.26}$$

where $|\mathbf{f}'(\cdot)|$ represents any matrix norm. Because $\mathbf{f}$ is contractive along the line, we can choose a matrix norm that is strictly less than 1. Thus,

$$|\mathbf{f}(\mathbf{m}^*) - \mathbf{f}(\mathbf{m})| < \int_0^1 |\mathbf{m}^* - \mathbf{m}| \, dt = |\mathbf{m}^* - \mathbf{m}|, \tag{2.27}$$

which is a contradiction. Thus $\mathbf{m}^* = \mathbf{m}$ and the stable non-negative steady-state is unique. Furthermore, this shows that $\{\mathbf{f}^\ell(\mathbf{1})\}$ converges to $\mathbf{m}$. □

As a corollary, we find that any stable non-negative steady-state attains the maximum total opinion among all steady-states.

Corollary 14. *Let* $(N, J, \mathbf{h}, \beta)$ *describe a system for which there exists a stable non-negative steady-state* $\mathbf{m}$*, and let* $\mathbf{m}'$ *be another steady-state. Then* $\mathbf{m} \geqslant \mathbf{m}'$.

Proof. By Lemmas 12 and 13 we have

$$\mathbf{m} = \lim_{\ell \to \infty} \mathbf{f}^\ell(\mathbf{1}) \geqslant \lim_{\ell \to \infty} \mathbf{f}^\ell(\mathbf{m}') = \mathbf{m}', \tag{2.28}$$

for any steady-state $\mathbf{m}'$. □

The concavity of stable non-negative steady-states in h

We show for J strongly-connected that any stable non-negative steady-state is concave in $\mathbf{h}$.

Lemma 15. *Let* $(N, J, \mathbf{h}, \beta)$ *describe a system with a strongly-connected graph for which there exists a stable non-negative steady-state* $\mathbf{m}$*. Then* $\mathbf{m}$ *is concave in* $\mathbf{h}$*.*

Proof. We want to show that the Hessian of m_i with respect to $\mathbf{h}$ is negative semidefinite for all $i \in N$. The Hessian of m_i with respect to $\mathbf{h}$ is given by

$$C^{(i)}_{jk} \equiv \frac{\partial^2 m_i}{\partial h_j \partial h_k} = \frac{\partial}{\partial h_k} \chi^{MF}_{ij}. \tag{2.29}$$

After taking partials and rearranging we are left with

$$C^{(i)}_{jk} = -2\sum_{\ell\in N} {\chi^{MF}_{j\ell}}^{T}\left(\frac{m_\ell}{(1-m_\ell^2)^2}\chi^{MF}_{i\ell}\right)\chi^{MF}_{\ell k} = -\sum_{\ell\in N} {Z^{(i)}_{j\ell}}^{T} Z^{(i)}_{\ell k}, \tag{2.30}$$

where $Z^{(i)}_{kj} = \chi^{MF}_{jk}\sqrt{\frac{2m_j}{(1-m_j^2)^2}\chi^{MF}_{ij}}$. Since $m_j, \chi^{MF}_{ij} \geqslant 0$ for all $i, j \in N$, $Z^{(i)}$ is a real matrix. Thus $C^{(i)}$ is negative semidefinite for all $i \in N$. □

2.8.2.2 *Theorem 5*

We show that any stable non-negative steady-state $\mathbf{m}$ gives rise to a unique and stable branch of steady-states that is smooth and non-decreasing as $\mathbf{h}$ increases. We also show that if $\mathbf{m} > 0$, then $\mathbf{m}$ gives rise to a unique and stable branch of steady-states that is smooth and non-decreasing as β increases. We note that the first result is given by Lemma 11. To prove the second result, we first show that any stable positive steady-state is locally non-decreasing in β.

Lemma 16. *Let* $(N, J, \mathbf{h}, \beta)$ *describe any system for which there exists a stable positive steady-state* $\mathbf{m}$*. Then* $\mathbf{m}$ *is locally non-decreasing in* β*.*

Proof. We want to show $\frac{dm_i}{d\beta}$ is non-negative for all $i \in N$. By assumption, $\rho(\mathbf{f}'(\mathbf{m})) < 1$, allowing us to apply the implicit function theorem, giving

$$\begin{aligned}\frac{dm_i}{d\beta} &= \sum_{j\in N}\left(\delta_{ji} - \mathbf{f}'(\mathbf{m})_{ji}\right)^{-1}\frac{\partial f_j}{\partial\beta} \\ &= \sum_{j\in N}\left(\delta_{ji} - \mathbf{f}'(\mathbf{m})_{ji}\right)^{-1}\operatorname{sech}^2\left[\beta(J\mathbf{m}+\mathbf{h})\right]_j (J\mathbf{m}+\mathbf{h})_j\,.\end{aligned} \tag{2.31}$$

In vector form,

$$\frac{d\mathbf{m}}{d\beta} = (I - \mathbf{f}'(\mathbf{m}))^{-1} D(\mathbf{m})(J\mathbf{m}+\mathbf{h}). \tag{2.32}$$

Theorem 4.3 of (*216*) guarantees that the matrix $(I - \mathbf{f}'(\mathbf{m}))^{-1}$ is non-negative and $D(\mathbf{m})$ is also non-negative. Because $\mathbf{m} = \tanh[\beta(J\mathbf{m}+\mathbf{h})] > 0$, we have $J\mathbf{m}+\mathbf{h} > 0$, and hence Eq. (2.32) is non-negative. □

We now complete the proof of Theorem 5, showing that any stable positive steady-state gives rise to a unique and stable branch of steady-states as β increases.

Proof (Theorem 5). For contradiction, assume that increasing β causes $\mathbf{m}$ to lose stability. Because the network is strongly-connected, $\mathbf{f}'(\mathbf{m}) = \beta D(\mathbf{m})J$ is also strongly-connected. Thus, Perron-Frobenius guarantees that $\mathbf{f}'(\mathbf{m})$ has a simple largest eigenvalue equal to its spectral radius. When $\mathbf{m}$ loses stability, this simple eigenvalue crosses one from below. By the Crandall-Rabinowitz theorem (*162*) and the principle of exchange of

stability, the crossing of the simple eigenvalue gives rise to two new sable steady-states. However, Lemma 16 guarantees that $\mathbf{m}$ remains positive as we increase β, which necessitates that both of the new stable steady-states are also initially positive, contradicting Theorem 4. Thus, $\mathbf{m}$ cannot lose stability as β increases, and hence $\mathbf{m}$ gives rise to a unique and smooth branch of stable and positive steady-states. □

2.8.2.3 *Theorem 3*

We show that every system exhibits a steady-state and that the well-known pitchfork bifurcation structure for steady-states of the ferromagnetic MF Ising model on a lattice extends exactly to general (weighted, directed) strongly-connected graphs. In particular, for any strongly-connected graph J, there is a critical interaction strength $\beta_c = 1/\rho(J)$ below which there exists a unique and stable steady-state. For $\mathbf{h} = 0$, as β crosses β_c from below, two new stable steady-states appear, one with all-positive components and one with all-negative components.

Proof (Theorem 3). We first note that the existence of a steady-state is guaranteed for any system by applying Brouwer's fixed point theorem to $\mathbf{f}$. Furthermore, Lemma 10 establishes that for $\beta < 1/\rho(J)$, there is a unique and stable steady-state.

In the case $\mathbf{h} = 0$, any system has a steady-state at $\mathbf{m}^* = 0$, which we refer to as the *trivial steady-state*. Lemma 10 guarantees that $\mathbf{m}^*$ is stable and unique for $\beta < \beta_c$. The implicit function theorem guarantees that we can continue to write $\mathbf{m}^*$ uniquely as a function of β so long as $\rho(\mathbf{f}'(\mathbf{m}^*)) = \beta\rho(J) < 1$. If our network is strongly-connected, then the Perron-Frobenius theorem guarantees that as we increase β, an eigenvalue of $\mathbf{f}'$ will first cross 1 when $\beta = 1/\rho(J)$. Furthermore, the largest eigenvalue is simple, which, by the Crandall-Rabinowitz theorem (*162*), guarantees the appearance of two new steady-states. Furthermore, the new solutions locally lie in the subspace spanned by the eigenvector corresponding to the largest eigenvalue of $\mathbf{f}'$, which by the Perron-Frobenius theorem has all positive components. Thus, at $\beta = \beta_c$, a branch of steady-states appears, giving rise to an all-positive steady-state and an all-negative steady-state. By the principle of exchange of stability, the new steady-states adopt the stability of the trivial steady-state, while the trivial steady-state becomes unstable. As we continue to increase β, Theorem 5 guarantees that the positive (negative) steady-state remains positive (negative) and stable. □

2.8.2.4 *Theorem 6*

We conclude by showing for J strongly-connected that if $\mathbf{h} \geqslant 0$, then there exists a stable non-negative steady-state.

Proof (Theorem 6). We first consider $\mathbf{h} > 0$. Lemma 10 guarantees that for any $\beta < \beta_c$ there exists a unique and stable steady-state $\mathbf{m}$ and that iterative application of $\mathbf{f}$ to any $\mathbf{x} \in X$ converges to $\mathbf{m}$. For induction, choose $\mathbf{x} = \mathbf{1}$ and assume $\mathbf{f}^{\ell}(\mathbf{1}) > 0$. Then

$$\mathbf{f}^{\ell+1}(\mathbf{1}) = \tanh\left[\beta(J\mathbf{f}^{\ell}(\mathbf{1}) + \mathbf{h})\right] > 0. \tag{2.33}$$

Thus, $\mathbf{m} = \lim_{\ell\to\infty} \mathbf{f}^{\ell}(\mathbf{1}) > 0$ at β. By Theorem 5, the unique branch $\mathbf{m}(\beta)$ remains stable and positive for all $\beta' > \beta$. To complete the proof, we note that Theorem 3 covers the case $\mathbf{h} = 0$. □

3

INFLUENCE MAXIMIZATION WITH THERMAL NOISE

This chapter contains work from Lynn, Christopher W., and Daniel D. Lee. "Statistical mechanics of influence maximization with thermal noise." EPL (Europhysics Letters) 117.6 (2017): 66001.

Abstract

The problem of optimally distributing a budget of influence among individuals in a social network, known as influence maximization, has typically been studied in the context of contagion models and deterministic processes, which fail to capture stochastic interactions inherent in real-world settings. Here, we show that by introducing thermal noise into influence models, the dynamics exactly resemble spins in a heterogeneous Ising system. In this way, influence maximization in the presence of thermal noise has a natural physical interpretation as maximizing the magnetization of an Ising system given a budget of external magnetic field. Using this statistical mechanical formulation, we demonstrate analytically that for small external field budgets, the optimal influence solutions exhibit a highly non-trivial temperature dependence, focusing on high-degree hub nodes at high temperatures and on easily-influenced peripheral nodes at low temperatures. For the general problem, we present a projected gradient ascent algorithm that uses the magnetic susceptibility to calculate locally-optimal external field distributions. We apply our algorithm to synthetic and real-world networks, demonstrating that our analytic results generalize qualitatively. Our work establishes a fruitful connection with statistical mechanics and demonstrates that influence maximization depends crucially on the temperature of the system, a fact that has not been appreciated by existing research.

3.1 INTRODUCTION

With the proliferation of online social networks, influence maximization has garnered tremendous attention as one of the paradigmatic problems towards the control of large complex networks (*190*, *475*, *556*), with applications ranging from marketing in social networks to immunization against infectious diseases. Given a network of social interactions and a budget of external influence, the goal of influence maximization is to distribute the budget among the nodes so as to maximize the total effect on the network. The problem was originally proposed in the context of viral marketing, and has since been studied primarily in the context of deterministic viral models (*265*, *275*,

365, *458*). However, these deterministic models neglect the important role of noise that is ubiquitous in real-world settings, such as the formation of opinions (*332*, *430*, *462*, *585*), the proliferation of technology (*454*), and the behavior of bounded-rational agents in games (*92*, *437*).

In this letter, we investigate the role of thermal noise in influence maximization. By injecting noise into the commonly used linear threshold model from sociology, we show that the ensuing dynamics are formally equivalent to Glauber Monte Carlo dynamics, simulating spins in an Ising system. In this way, we show that influence maximization with thermal noise has a natural physical interpretation as maximizing the magnetization of a heterogeneous Ising system given a budget of external magnetic field (*415*).

We find that the structure of solutions has a highly non-trivial dependence on the temperature of the system. For small budgets, we demonstrate analytically that at high temperatures, the optimal external field distribution focuses on hub nodes with large degrees. On the other hand, at low temperatures, because hub nodes are strongly bound in the ground state, we show that it is optimal to focus the external field on peripheral nodes that are easily influenced. In addition to our analysis, we also present a projected gradient ascent algorithm that uses the magnetic susceptibility to efficiently calculate locally optimal solutions for general influence maximization problems. Using Monte Carlo simulations to estimate the susceptibility at each step, we apply our algorithm on large real-world networks, showing that the structure of influence maximization solutions is qualitatively similar to our analytic description. Together, our results show that influence maximization depends crucially on the temperature of the system, a finding that can lead to numerous practical consequences.

3.2 GLAUBER DYNAMICS

We introduce noise into a commonly-used influence model from sociology known as the linear threshold model, and show that the resulting dynamics exactly resemble spins in an Ising system. The Ising model has previously been proposed to describe social interactions, most notably in (*238*, *239*). The novel contribution of this section is to draw a formal equivalence between the linear threshold model with noise and the Ising model, allowing us to leverage statistical mechanical tools to study influence maximization with thermal noise.

The linear threshold model has found wide use in the influence maximization literature (*364*, *459*), and is closely related to the independent cascade model (*365*), best-response dynamics in game theory (*234*), and percolation in complex networks (*458*). We consider a social network consisting of a set of n nodes $\{\sigma_i, i = 1, \ldots, n\}$, each of which is either active ($\sigma_i = +1$) or inactive ($\sigma_i = -1$). The connections in the social network are described by a coupling matrix $J \in \mathbb{R}^{n \times n}$, where J_{ij} represents the influence that node j holds over node i. Each node i has an associated threshold $\theta_i \in \mathbb{R}$, which represents the total weight of its neighbors needed for the node to become active.

At each time step t, a node i is chosen at random and its activity is updated according to the following rule:

$$\sigma_i^{(t+1)} = \text{sign}\left[\sum_{j\neq i} J_{ij}\sigma_j^{(t)} - \theta_i\right]. \tag{3.1}$$

The dynamics in (4.12) are deterministic; in order to model stochastic interactions inherent to real-world settings, we assume that the individual being updated at time t experiences an additional random influence $\epsilon^{(t)}$, such that the linear threshold dynamics become

$$\sigma_i^{(t+1)} = \text{sign}\left[\sum_{j} J_{ij}\sigma_j^{(t)} - \theta_i + \epsilon^{(t)}\right]. \tag{3.2}$$

If $\epsilon^{(t)}$ are drawn i.i.d. from a logistic distribution $p(\epsilon)$ with mean zero and variance $T^2\pi^2/12$, then the dynamics in (3.2) are equivalent to the following probabilistic update rule:

$$\begin{aligned} P\left(\sigma_i^{(t+1)} = 1|\boldsymbol{\sigma}^{(t)}\right) &= \int_{-\left(\sum_j J_{ij}\sigma_j^{(t)} - \theta_i\right)}^{\infty} p(\epsilon)d\epsilon \\ &= \frac{1}{1+\exp\left[-\frac{2}{T}\left(\sum_j J_{ij}\sigma_j^{(t)} - \theta_i\right)\right]}. \end{aligned} \tag{3.3}$$

This stochastic process is formally equivalent to Glauber dynamics describing Ising systems (*399*). Thus, the linear threshold model with logistically-distributed noise can be understood as an Ising model with temperature T, exchange couplings J, and a heterogeneous external field $b_i \equiv -\theta_i$.

In this letter, we assume J is symmetric so that the Glauber dynamics settle to thermal equilibrium described by the Boltzmann distribution:

$$P(\boldsymbol{\sigma}) = \frac{1}{Z}e^{-\frac{1}{T}H(\boldsymbol{\sigma})}, \tag{3.4}$$

with Hamiltonian

$$H(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i b_i\sigma_i, \tag{3.5}$$

and partition function

$$Z = \sum_{\{\boldsymbol{\sigma}\}} e^{-\frac{1}{T}H(\boldsymbol{\sigma})}. \tag{3.6}$$

While assuming $J = J^T$ is reasonable in some networks (e.g., friendships on Facebook), we note that there are settings where the network may be far from symmetric (e.g., followers on Twitter), in which case the Glauber dynamics do not admit a steady-state description. Thus, we use the Boltzmann distribution as an important first step

toward understanding the equilibrium properties of influence maximization, noting that further research should investigate cases where J is not symmetric.

Using this statistical mechanical formulation, we can represent important equilibrium quantities as expectations over the Boltzmann distribution, denoted by $\langle\cdot\rangle$. For example, the total number of expected activations has a physical interpretation as the total magnetization under applied field $\mathbf{b}$:

$$M(\mathbf{b}) = \sum_i \langle \sigma_i \rangle , \tag{3.7}$$

Because we are interested in tuning the magnetization with respect to external influence, another important quantity is the susceptibility vector, defined by

$$\chi_i = \frac{\partial M(\mathbf{b})}{\partial b_i} = \frac{1}{T} \sum_j \left(\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \right) . \tag{3.8}$$

Besides the symmetry of J, we note that our main analysis does not rely on additional assumptions about the signs of the couplings or external fields. Thus, while previous work has only considered positive couplings (*415*), in principle both ferromagnetic ($J_{ij} > 0$) and antiferromagnetic ($J_{ij} < 0$) bonds can be allowed. In the presence of significant frustration, our model becomes similar to a spin-glass, which is known to exhibit much more complex behavior than a standard Ising ferromagnet (*483*). In our numerical examples, however, we will show influence maximization results on ferromagnetic systems in order to ease the discussion.

3.3 ISING INFLUENCE MAXIMIZATION

Influence maximization, as originally defined in (*365*), is the problem of choosing a subset of nodes to initially activate such that the ensuing spread of activations is maximized. This problem has been shown to be computationally hard for a generic class of linear threshold models, including Eq. (4.12) (*365*, *459*). Indeed, identifying the optimal set of seed activations is a combinatorial optimization problem involving many-body interactions, where the topology of the social network plays a crucial role.

Here we consider how the total magnetization increases after the addition of an external field $\mathbf{h}$, which is bounded by a budget H. If $\mathbf{b}^0$ is the initial external field, then the total applied field is $\mathbf{b} = \mathbf{b}^0 + \mathbf{h}$. Incrementing the field h_i at node i will increase the expected local activation $\langle \sigma_i \rangle$, but it is the indirect effects of the external field mediated by the network connections that give influence maximization solutions a rich structure.

Specifically, we study the problem of maximizing the magnetization of an Ising system with respect to an additional external field $\mathbf{h}$, subject to the budget constraint $|\mathbf{h}|_p = (\sum_i |h_i|^p)^{1/p} \leqslant H$, where $H > 0$ is the external field budget, and we denote the set of feasible external fields by $\mathcal{F}_H = \{\mathbf{h} \in \mathbb{R}^n : |\mathbf{h}|_p \leqslant H\}$. Thus, given an initial

Ising system described by $J = J^T$, $\mathbf{b}^0$, and T, and an external field budget H, the Ising influence maximization problem is to find an optimal external field satisfying

$$\mathbf{h}^* = \arg\max_{\mathbf{h} \in \mathcal{F}_H} M(\mathbf{b}^0 + \mathbf{h}). \tag{3.9}$$

We remark that when $p = 0$, the ℓ_0-norm budget constraint is equivalent to the original discrete constraint in (*365*). Hence, this formulation can be viewed as an equilibrium version of the traditional influence maximization problem with non-zero temperature.

3.4 SMALL H BUDGET

The Ising influence maximization (IIM) problem in Eq. (4.2) was first studied using the naïve mean-field approximation on ferromagnetic systems in (*415*). In this section, we consider a general class of Ising systems and their solutions in the limit of small budget H. In this limit, any feasible external field $\mathbf{h} \in \mathcal{F}_H$ will induce a small change in the magnetization that can be approximated by the linear response relation

$$\Delta M = M(\mathbf{b}^0 + \mathbf{h}) - M(\mathbf{b}^0) \approx \boldsymbol{\chi}(\mathbf{b}^0)^T \mathbf{h}. \tag{3.10}$$

Thus, in the small-H limit, the optimal external field is a function only of the magnetic susceptibility, focusing on nodes that correspond to larger entries in $\boldsymbol{\chi}$. For instance, under an ℓ_2 budget constraint, the optimal external field points in the direction of the susceptibility vector:

$$\mathbf{h}^* = H\hat{\boldsymbol{\chi}}, \tag{3.11}$$

where $\hat{\boldsymbol{\chi}} = \boldsymbol{\chi}/|\boldsymbol{\chi}|_2$. On the other hand, under an ℓ_1 budget constraint, assuming $\boldsymbol{\chi}$ has a unique largest entry, the optimal external field focuses on the node with the largest susceptibility:

$$h_i^* = \begin{cases} H, & \text{if } i = \arg\max_j \chi_j \\ 0, & \text{otherwise} \end{cases}. \tag{3.12}$$

Since, in the small-H limit, the optimal external field focuses on the nodes with large susceptibilities, we can gain an intuition for the temperature dependence of influence maximization by considering how the structure of $\boldsymbol{\chi}$ depends on T. In Fig. 3.1, we consider a small ferromagnetic network in a uniform positive external field, and we note that the spectral radius $\rho(J)$ sets the temperature scale. At high temperatures, we find that the largest component of the susceptibility intuitively corresponds to the hub node of largest degree. As the temperature decreases, the susceptibility exhibits an abrupt shift such that, at low temperatures, the largest components of $\boldsymbol{\chi}$ counter-intuitively correspond to peripheral nodes of low degree. As evidenced by this example, introducing thermal noise into influence maximization can induce a highly non-trivial dependence in the solution structure that is not observed under traditional deterministic models.

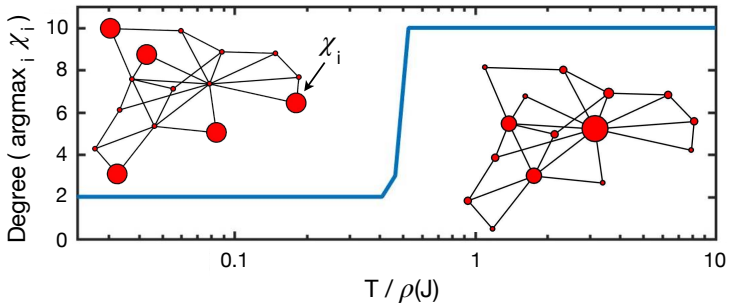


Figure 3.1: Shift in the structure of the susceptibility. We consider a small ferromagnetic network with $J_{ij} = J_{ji} \in \{0, 1\}$ and a uniform positive external field $\mathbf{b}^0 = 0.3$. At high temperatures, the node corresponding to the largest entry in $\boldsymbol{\chi}$ is the hub node of degree 10, while, at low temperatures, the nodes with the largest susceptibilities are the peripheral nodes of degree 2. Thus, for small H, the optimal external field shifts from focusing on the hub node at high temperatures to the low-degree nodes at low temperatures. The magnitudes of the entries in $\boldsymbol{\chi}$ are represented by the sizes of the nodes in the network snapshots.

To understand the significance of thermal noise in influence maximization, we begin by developing an analytic description of small-budget optimal external fields in the high- and low-temperature limits. In the high-T limit, we demonstrate that the optimal external fields focus on hub nodes of high degree because the effects of the external field are localized. On the contrary, in the low-temperature limit, we show that the optimal external field focuses on easily-influenced nodes which are minimally energetically bound to the ground state. Because flipping a hub node from the ground state often incurs a large energetic cost, these easily-influenced nodes tend to have low degree.

3.4.1 *High-temperature solution*

In the high-temperature limit, we can expand $\boldsymbol{\chi}$ in powers of $\beta \equiv \frac{1}{T}$:

$$\chi_i = \chi_i|_{\beta=0} + \beta \frac{\partial}{\partial \beta} \chi_i \bigg|_{\beta=0} + \frac{\beta^2}{2} \frac{\partial^2}{\partial \beta^2} \chi_i \bigg|_{\beta=0} + \dots . \tag{3.13}$$

For a general Ising system defined by J, $\mathbf{b}$, and T, we calculate the terms in (3.21) up to third-order in β (see Sec. **??**), yielding

$$\chi_i = \beta + \beta^2 d_i + \beta^3 \left(\sum_{j \neq i} (J^2)_{ij} - b_i^2 \right) + \dots , \tag{3.14}$$

where we define $d_i = \sum_j J_{ij}$ as the degree of node i and $\sum_{j \neq i} (J^2)_{ij} = \sum_{jk} J_{ik} J_{kj} - \sum_j J_{ij}^2$ as the second-degree of node i, representing the combined weight of paths of length 2 originating from node i (excluding self-interactions).

Inspecting Eq. (3.14), we find to first-order in β that $\chi_i > 0$. This indicates that the high-T optimal external fields $\mathbf{h}^*$ are positive. This may seem obvious, but we note

that at intermediate and low temperatures, applying a negative external field to nodes that are anti-correlated with the total magnetization can be optimal. Furthermore, since the second-order term in Eq. (3.14) is proportional to d_i, the high-temperature optimal external fields focus H on the nodes of high degree, independently of the initial external field $\mathbf{b}^0$. Since the external field only appears in Eq. (3.14) to third-order in β, this high-temperature argument generalizes to non-trivial budgets, so long as $\beta^3 H^2 \ll 1$. Finally, we remark that focusing one's budget on the nodes of high degree is intuitive and has often been used as a heuristic for influence maximization problems.

3.4.2 *Low-temperature solution*

In the low-temperature limit, the susceptibility is dominated by the structure of low-energy states, and we find that the nodes with the largest susceptibilities are those with opposite parity between the ground and first-excited states. In ferromagnetic systems, we find that these nodes tend to have low degree in the network, in stark contrast to the high-temperature susceptibility.

Here we assume that the system admits a unique ground state $\boldsymbol{\sigma}^0$ and a unique first-excited state $\boldsymbol{\sigma}^1$, and we consider more general cases in the SM. Letting $\Delta E = H(\boldsymbol{\sigma}^1) - H(\boldsymbol{\sigma}^0) > 0$ denote the energy gap between the ground and first-excited states, at low temperatures, the node magnetizations can be approximated by:

$$\langle \sigma_i \rangle \approx \frac{\sigma_i^0 + \sigma_i^1 e^{-\frac{\Delta E}{T}}}{1 + e^{-\frac{\Delta E}{T}}} \approx \sigma_i^0 + \left(\sigma_i^1 - \sigma_i^0\right) e^{-\frac{\Delta E}{T}}. \tag{3.15}$$

Similarly, the two-point correlations can be approximated by:

$$\left\langle \sigma_i \sigma_j \right\rangle \approx \sigma_i^0 \sigma_j^0 + \left(\sigma_i^1 \sigma_j^1 - \sigma_i^0 \sigma_j^0\right) e^{-\frac{\Delta E}{T}}. \tag{3.16}$$

Using Eq. (3.8), we can write the low-temperature susceptibility in terms of the approximate one- and two-point correlations, yielding

$$\begin{aligned}
\chi_i &= \frac{1}{T} \sum_j \left(\left\langle \sigma_i \sigma_j \right\rangle - \langle \sigma_i \rangle \left\langle \sigma_j \right\rangle \right) \\
&\approx \frac{1}{T} \sum_j \Big[\left(\sigma_i^0 \sigma_j^0 + \left(\sigma_i^1 \sigma_j^1 - \sigma_i^0 \sigma_j^0\right) e^{-\frac{\Delta E}{T}} \right) \\
&\qquad - \left(\sigma_i^0 + \left(\sigma_i^1 - \sigma_i^0\right) e^{-\frac{\Delta E}{T}} \right) \left(\sigma_j^0 + \left(\sigma_j^1 - \sigma_j^0\right) e^{-\frac{\Delta E}{T}} \right) \Big] \\
&\approx \frac{e^{-\frac{\Delta E}{T}}}{T} \left(\sigma_i^1 - \sigma_i^0\right) \sum_j \left(\sigma_j^1 - \sigma_j^0\right),
\end{aligned} \tag{3.17}$$

where the final approximation follows by keeping only those terms of order $e^{-\frac{\Delta E}{T}}/T$, since all other terms are exponentially suppressed for small T. Eq. (3.17) reveals insight into the structure of low-temperature optimal external fields for systems with

unique ground and first-excited states. Since χ_i is proportional to $(\sigma_i^1 - \sigma_i^0)$, the low-T susceptibility is only non-zero for nodes with opposite parity between the ground and first-excited states, i.e., nodes i for which $\sigma_i^1 \neq \sigma_i^0$.

If only one node flips between the ground and first-excited state, then this node can be thought of as "easily-influenced" in the sense that it induces a minimal increase in energy when flipped from the ground state. If, in addition, the system is ferromagnetic ($J \geqslant 0$) in a uniform external field, then this node must have the lowest degree in the network. We remark that much of the intuition developed here extends to general systems and budgets H of arbitrary size (see SM). Finally, we emphasize the stark dissimilarity between the high- and low-T susceptibilities, highlighting the important role that thermal noise plays in determining the structure of optimal external fields.

3.5 EXACT SOLUTIONS FOR SMALL H BUDGET

Thus far, we have established that the structure of optimal external fields has a highly non-trivial dependence on the temperature of the system. To make this point concrete, we consider a family of heterogeneous one-dimensional systems for which the susceptibility can be efficiently calculated using transfer matrices (**Chs24-Kramers-01**). Specifically, we study a 1-D ring of nodes $i \in \{1, \ldots, n\}$, where node i is coupled to node $i+1$ with weight

$$J_{i,i+1} = \frac{1}{2}\left[\cos\left(\frac{\theta_i + \theta_{i+1}}{2}\right) + 1\right], \tag{3.18}$$

where $\theta_i = 2\pi(i-1)/n$ is the angle of the node i around the ring (Fig. 3.2a). Node 1 ($\theta = 0$) has the largest degree in the network and node $n/2+1$ ($\theta = \pi$) has the smallest degree. The system is placed in a uniform external field $\mathbf{b} = 0.1$ such that the ground-state configuration is "all-up" and the first-excited state corresponds to the node at $\theta = \pi$ flipping down. Thus, at high temperatures, we expect $\chi(\theta)$ to have a maximum at $\theta = 0$, and as the temperature decreases, the maximum should shift to $\theta = \pi$.

Fig. 3.2b shows the temperature-dependence of the susceptibility for a ring of 100 nodes. At high temperatures, the susceptibility is nearly uniform with the node at $\theta = 0$ having the largest entry, while at low temperatures, the susceptibility becomes localized around $\theta = \pi$. Thus, in addition to shifting from low- to high-degree nodes, the susceptibility also changes from being nearly uniform to localized as the temperature decreases. Furthermore, at intermediate temperatures, nodes other than those of the highest or lowest degrees (e.g., $\theta = \pi/2$) can have the largest susceptibility.

Fig. 3.2c shows the full structure of the susceptibility as a function of θ at three different temperatures and for a number of system sizes. The right and left panels clearly demonstrate the shift from uniformity to localization of the susceptibility, respectively. Furthermore, as the systems become larger, $n \to \infty$, the susceptibility converges to a well-defined structure that maintains a non-trivial dependence on

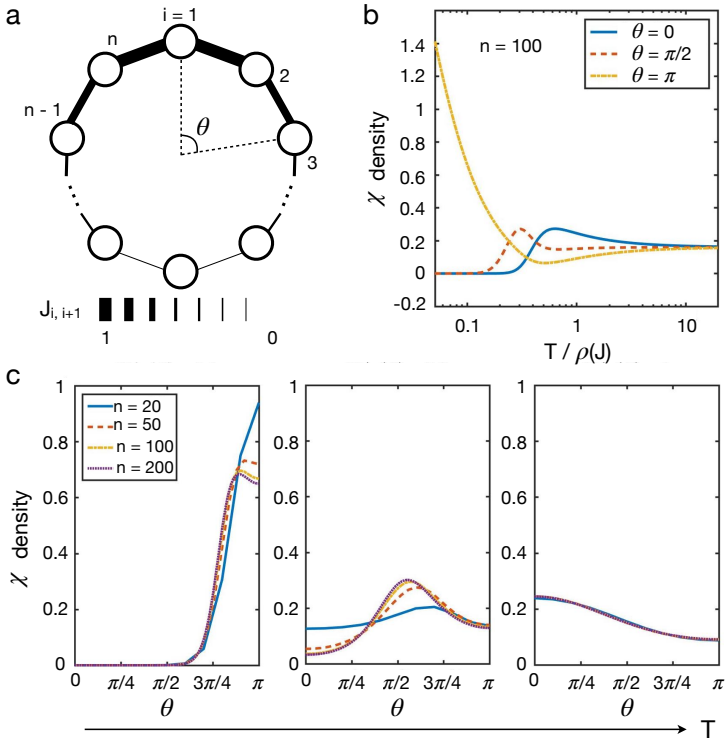


Figure 3.2: Temperature-dependence of the susceptibility in a heterogeneous ring. (*a*) The ring has nearest-neighbor couplings $J_{i,i+1}$ defined in Eq. (3.18) and a positive uniform external field. (*b*) At high temperatures, χ is nearly uniform and the largest entry corresponds to the node of highest degree ($\theta = 0$). At low temperatures, the susceptibility is localized near the node of lowest degree ($\theta = \pi$). (*c*) The susceptibility density is normalized such that the integral over all angles is unity, and is shown as a function of the angle for various temperatures T and system sizes n.

T. We also studied networks with community structure, such as those in (*105*), and found that the structure of the susceptibility differed greatly between the high- and low-temperature limits.

3.6 NUMERICAL TECHNIQUES FOR GENERAL ISING SYSTEMS

In the small budget limit, the optimal external field is determined by the susceptibility of the initial system. For larger external field budgets H, however, one must take into account how the magnetization varies as $\mathbf{h}$ increases. Directly trying to search for the maximum of M with respect to $\mathbf{h} \in \mathcal{F}_H$, which has n components, is computationally intractable for larger systems. To overcome this problem, we present a gradient ascent algorithm that iteratively calculates the susceptibility and efficiently converges to a local maximum of the magnetization. While previous work has focused on maximizing the mean-field magnetization (*415*), our algorithm maximizes the exact magnetization for general Ising systems and general external field budgets.

The algorithm is first initialized at a feasible external field $\mathbf{h}^{(0)} \in \mathcal{F}_H$ and steps along the direction of the gradient of the objective, given by the magnetic susceptibility $\boldsymbol{\chi}(\mathbf{b}) = \nabla M(\mathbf{b})$. Projecting back onto the space of feasible external fields completes each iteration:

$$\mathbf{h}^{(k+1)} \leftarrow P_{\mathcal{F}_H}\left[\mathbf{h}^{(k)} + \alpha\boldsymbol{\chi}(\mathbf{b}^0 + \mathbf{h}^{(k)})\right], \tag{3.19}$$

where $P_{\mathcal{F}_H}$ denotes the projection onto $\mathcal{F}_H$, which is convex for norms $p \geqslant 1$, and $\alpha \geqslant 0$ is the step size. If $\boldsymbol{\chi}$ is computed exactly at each step, then this algorithm converges to an ϵ-approximate local maximum of M in $O(1/\epsilon)$ iterations (see (*471*)). Detailed pseudocode for our algorithm is presented in the SM.

We apply our algorithm on random and real-world networks to test its performance and to probe the structure of optimal external fields for larger budgets H. Although our algorithm is efficient and converges in relatively few iterations, the total run-time is limited by the calculation of $\boldsymbol{\chi}$ at each step. Exact calculations for generic Ising systems are limited to relatively small networks. Thus, to scale our algorithm to larger systems, we can approximate $\boldsymbol{\chi}$ using Monte Carlo (MC) techniques at each iteration.

In Fig. 3.3, we show the results of exact and MC implementations of our algorithm, with solutions denoted by $\mathbf{h}^*$ and $\mathbf{h}^*_{MC}$, along the mean-field solution in (*415*), denoted $\mathbf{h}^*_{MF}$, for a small ferromagnetic random network. With an ℓ_1 budget constraint, Fig. 3.3a shows that all three algorithms shift from focusing on the node of highest degree at high temperatures to the nodes of low degree at low temperatures. Fig. 3.3b compares the magnetizations achieved by the three algorithms, along with the magnetization achieved by a uniform external field $\mathbf{h}_{\mathtt{uniform}} = \frac{H}{n}(1,\ldots,1)^T$ as a baseline. As $T \to 0$, thermal noise dominates the external fields and all strategies yield $M = 0$. As $T \to 0$, any choice of positive external field yields $M = n$ as the system approaches the all-up ground state. At intermediate temperatures, however, Fig. 3.3b verifies that the exact implementation of our algorithm achieves the largest magnetization among the algorithms considered.

In Fig. 3.4, we apply the MC implementation of our algorithm on a real-world network of co-authorships on arXiv (*395*, *477*). The co-authorship network consists of $n = 904$ nodes, where each node represents a physicist, and each edge represents the co-authorship of a paper on arXiv. For an ℓ_1 budget constraint, Fig. 3.4a illustrates the shifts in $\mathbf{h}^*_{MC}$ and $\mathbf{h}^*_{MF}$ from focusing on high- to low-degree nodes as the temperature decreases. Thus, in the networks presented here, the optimal external fields for non-trivial budgets closely resemble our small-budget descriptions. We have found the same structure in other large random networks as well, which have been omitted to save space. Fig. 3.4b compares the performances of $\mathbf{h}^*_{MC}$ and $\mathbf{h}^*_{MF}$, where each data point represents the average magnetization from 20 Monte Carlo simulations. This demonstrates that our algorithm scales to large systems and performs favorably in comparison to the mean-field algorithm.

We emphasize the implications of our analysis and numerical results. By including thermal noise in influence maximization, the structure of solutions acquires a highly non-trivial dependence on the temperature of the system. Thus, in the control of noisy

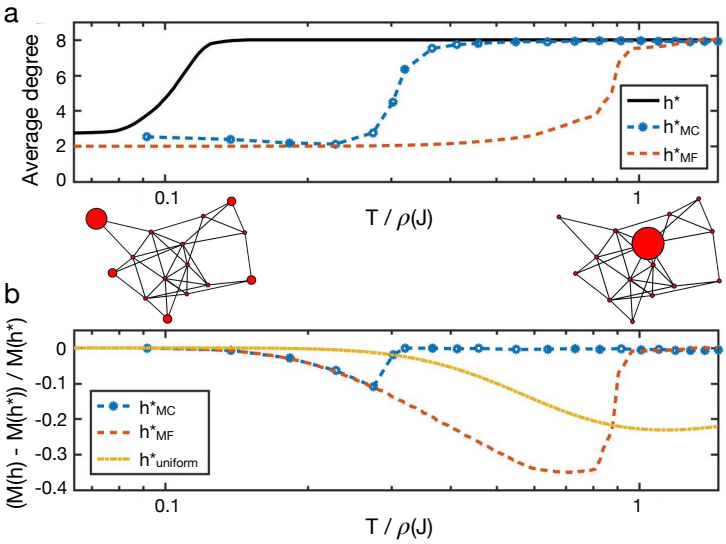


Figure 3.3: Shift in solution structure for a small Erdös-Rényi network. (*a*) We consider an Erdös-Rényi network with $n = 15$ nodes, $J_{ij} = J_{ji} \in \{0, 1\}$, and $\mathbf{b}^0 = 0$. For $H = 1$ and for an ℓ_1 constraint, we find that $\mathbf{h}^*$, $\mathbf{h}^*_{MC}$, and $\mathbf{h}^*_{MF}$ all shift from focusing on high- to low-degree nodes as T decreases. The network snapshots illustrate the allocations of the budget in the high- and low-temperature limits. (*b*) We compare $M(\mathbf{h}^*)$ with the magnetizations under $\mathbf{h}^*_{MC}$, $\mathbf{h}^*_{MF}$, and $\mathbf{h}_{uniform}$, verifying that $\mathbf{h}^*$ achieves the highest magnetization across all temperatures and that $\mathbf{h}^*_{MC}$ compares favorably.

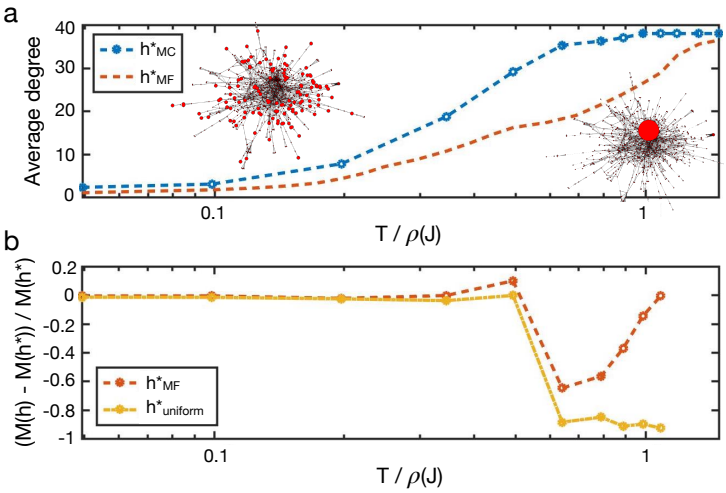


Figure 3.4: Shift in solution structure for a real-world social network. (*a*) We consider a co-authorship network with $n = 904$ nodes, $J_{ij} = J_{ji} \in \{0, 1\}$, and $\mathbf{b}^0 = 0$.For an ℓ_1 budget constraint with $H = 20$, $\mathbf{h}^*_{MC}$ and $\mathbf{h}^*_{MF}$ both shift from focusing on high- to low-degree nodes, illustrated by the network snapshots. (*b*) We compare $M(\mathbf{h}^*_{MC})$ with the magnetizations under $\mathbf{h}^*_{MF}$ and $\mathbf{h}_{uniform}$, demonstrating that $\mathbf{h}^*_{MC}$ achieves the highest magnetization across most temperatures.

systems, accurately accounting for the strength of random fluctuations is critically

important. Because random fluctuations are ubiquitous in nature, the important role of thermal noise in real-world applications should not be underestimated.

3.7 CONCLUSIONS

We study influence maximization in the presence of thermal fluctuations. By introducing thermal noise into the commonly-used linear threshold model, we show that the dynamics are equivalent to Glauber dynamics for an Ising system. In this way, influence maximization with thermal noise has a natural physical interpretation as maximizing the magnetization given a budget of external magnetic field. In the limit of small budget, we demonstrate that the structure of solutions given by the magnetic susceptibility exhibits a highly non-trivial temperature dependence, focusing on high-degree hub nodes at high temperatures, while focusing on easily-influenced peripheral nodes at low temperatures. For general systems and budgets, we present a projected gradient ascent algorithm that iteratively calculates the susceptibility and efficiently converges to local maxima of the magnetization. In a number of random and real-world networks, we demonstrate that our numerical results can be qualitatively understood using our analysis. Our work establishes fruitful connections between statistical physics, machine learning, and network science, paving the way for future cross-disciplinary research.

3.8 SUPPLEMENTARY MATERIAL

3.8.1 *High-temperature susceptibility*

For a general Ising system described by the coupling matrix $J = J^T \in \mathbb{R}^{n\times n}$ (where we assume $J_{ii} = 0$ for all $i \in N$), heterogeneous external field $\mathbf{b} \in \mathbb{R}^n$, and temperature $T > 0$, we consider the susceptibility vector defined by:

$$\chi_i = \sum_j \frac{\partial \langle \sigma_j \rangle}{\partial b_i} = \frac{1}{T} \sum_j \left(\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle \right), \tag{3.20}$$

where the sums run over all nodes $i \in \{1, \dots, n\}$ and $\langle \cdot \rangle$ denotes an expectation over the Boltzmann distribution. The component χ_i quantifies the response of the total magnetization $M = \sum_j \langle \sigma_j \rangle$ to a change in the external field on node i. For high temperatures, we can expand the susceptibility in powers of $\beta \equiv \frac{1}{T}$,

$$\frac{1}{\beta}\chi_i = \left(\frac{1}{\beta}\chi_i\right)_{\beta=0} + \beta \frac{\partial}{\partial \beta}\left(\frac{1}{\beta}\chi_i\right)_{\beta=0} + \dots. \tag{3.21}$$

In what follows, we derive the susceptibility up to order β^3.

First, we must establish some preliminary results. Since the Boltzmann distribution is uniform at $\beta = 0$, for any function of the spins $f(\boldsymbol{\sigma})$ we have

$$\langle f(\boldsymbol{\sigma}) \rangle_{\beta=0} = \frac{1}{2^n} \sum_{\{\boldsymbol{\sigma}\}} f(\boldsymbol{\sigma}), \tag{3.22}$$

where the sum runs over the set of spin configurations $\{\pm 1\}^n$. Thus, when evaluated at $\beta = 0$, any terms in $f(\boldsymbol{\sigma})$ that involve an odd number of spins will vanish in expectation. For terms involving even numbers of spins, we have

$$\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}} \sigma_i \sigma_j = \delta_{ij}, \tag{3.23}$$

$$\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}} \sigma_i \sigma_j \sigma_k \sigma_\ell = \delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk} - 2\delta_{ijk\ell} \triangleq C_{ijk\ell}, \tag{3.24}$$

$$\begin{aligned}
\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}} \sigma_i \sigma_j \sigma_k \sigma_\ell \sigma_m \sigma_p =& \delta_{ij}(\delta_{k\ell}\delta_{mp} + \delta_{km}\delta_{\ell p} + \delta_{kp}\delta_{\ell m}) \\
&+ \delta_{ik}(\delta_{j\ell}\delta_{mp} + \delta_{jm}\delta_{\ell p} + \delta_{jp}\delta_{\ell m}) \\
&+ \delta_{i\ell}(\delta_{jk}\delta_{mp} + \delta_{jm}\delta_{kp} + \delta_{jp}\delta_{km}) \\
&+ \delta_{im}(\delta_{jk}\delta_{\ell p} + \delta_{j\ell}\delta_{kp} + \delta_{jp}\delta_{k\ell}) \\
&+ \delta_{ip}(\delta_{jk}\delta_{\ell m} + \delta_{j\ell}\delta_{km} + \delta_{jm}\delta_{k\ell}) \\
&- 2\left[\delta_{ij}\delta_{k\ell mp} + \delta_{ik}\delta_{j\ell mp} + \delta_{i\ell}\delta_{jkmp}\right. \\
&+ \delta_{im}\delta_{jk\ell p} + \delta_{ip}\delta_{jk\ell m} + \delta_{jk}\delta_{i\ell mp} \\
&+ \delta_{j\ell}\delta_{ikmp} + \delta_{jm}\delta_{ik\ell p} + \delta_{jp}\delta_{ik\ell m} \\
&+ \delta_{k\ell}\delta_{ijmp} + \delta_{km}\delta_{ij\ell p} + \delta_{kp}\delta_{ij\ell m} \\
&\left.+\delta_{\ell m}\delta_{ijkp} + \delta_{\ell p}\delta_{ijkm} + \delta_{mp}\delta_{ijk\ell}\right] \\
&+ 16\delta_{ijk\ell mp} \\
&\triangleq C_{ijk\ell mp},
\end{aligned} \tag{3.25}$$

where δ_{ij} denotes the Kronecker delta. Furthermore, we note that the derivative of the expectation of any function $f(\boldsymbol{\sigma})$ with respect to β takes the form

$$\frac{\partial}{\partial \beta}\langle f(\boldsymbol{\sigma})\rangle = -\langle f(\boldsymbol{\sigma}) H(\boldsymbol{\sigma})\rangle + \langle f(\boldsymbol{\sigma})\rangle \langle H(\boldsymbol{\sigma})\rangle, \tag{3.26}$$

where $H(\boldsymbol{\sigma}) = -\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j - \sum_i b_i\sigma_i$ is the Hamiltonian of the system.

We proceed with the zeroth-order term in Eq. (3.21), which takes the form

$$\begin{aligned}
\left(\frac{1}{\beta}\chi_i\right)_{\beta=0} &= \sum_j \left(\langle \sigma_i\sigma_j\rangle_{\beta=0} - \cancel{\langle \sigma_i\rangle_{\beta=0}}\cancel{\langle \sigma_j\rangle_{\beta=0}}\right) \\
&= \sum_j \left(\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}} \sigma_i\sigma_j\right) = \sum_j \delta_{ij} = 1.
\end{aligned} \tag{3.27}$$

Before calculating the first-order term in Eq. (3.21), we note that

$$\langle H(\boldsymbol{\sigma})\rangle_{\beta=0} = -\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}}\left(\frac{1}{2}\sum_{ij} J_{ij}\sigma_i\sigma_j + \sum_i b_i\sigma_i\right) = -\frac{1}{2}\sum_{ij} J_{ij}\delta_{ij} = 0, \tag{3.28}$$

since $J_{ii} = 0$ for all $i \in N$. Thus, we have

$$\begin{aligned}
\frac{\partial}{\partial\beta}\left(\frac{1}{\beta}\chi_i\right)_{\beta=0} &= \sum_j \Big[-\langle\sigma_i\sigma_j H(\boldsymbol{\sigma})\rangle + \cancel{\langle\sigma_i\sigma_j\rangle\langle H(\boldsymbol{\sigma})\rangle} \\
&\qquad \cancel{-\langle\sigma_i\rangle\left(-\langle\sigma_j H(\boldsymbol{\sigma})\rangle + \langle\sigma_j\rangle\langle H(\boldsymbol{\sigma})\rangle\right)} \\
&\qquad \cancel{-\langle\sigma_j\rangle\left(-\langle\sigma_i H(\boldsymbol{\sigma})\rangle + \langle\sigma_i\rangle\langle H(\boldsymbol{\sigma})\rangle\right)}\Big]_{\beta=0} \\
&= \sum_j \left[\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}}\sigma_i\sigma_j\left(\frac{1}{2}\sum_{k\ell} J_{k\ell}\sigma_k\sigma_\ell + \sum_k b_k\sigma_i\right)\right] \\
&= \frac{1}{2}\sum_{jk\ell} J_{k\ell}\left(\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk} - 2\delta_{ijk\ell}\right) \\
&= \frac{1}{2}\sum_j \left(J_{ij} + J_{ji}\right) = \sum_j J_{ij},
\end{aligned} \tag{3.29}$$

where the penultimate equality follows from J being zero on the diagonal and the final equality follows since J is symmetric. We identify $\sum_j J_{ij} \equiv d_i$ as the degree of node i.

We now derive the second-order term in Eq. (3.21). Including only the non-vanishing terms, we have

$$\begin{aligned}
\frac{\partial^2}{\partial\beta^2}\left(\frac{1}{\beta}\chi_i\right)_{\beta=0} = \sum_j \Big[&\langle\sigma_i\sigma_j H^2(\boldsymbol{\sigma})\rangle - \langle\sigma_i\sigma_j\rangle\langle H^2(\boldsymbol{\sigma})\rangle \\
&-2\langle\sigma_i H(\boldsymbol{\sigma})\rangle\langle\sigma_j H(\boldsymbol{\sigma})\rangle\Big]_{\beta=0}.
\end{aligned} \tag{3.30}$$

Evaluating the second term in the sum in Eq. (3.30), we have

$$\begin{aligned}
\langle H^2(\boldsymbol{\sigma})\rangle_{\beta=0} &= \frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}}\left(\frac{1}{2}\sum_{k\ell} J_{k\ell}\sigma_k\sigma_\ell + \sum_k b_k\sigma_k\right)\left(\frac{1}{2}\sum_{mp} J_{mp}\sigma_m\sigma_p + \sum_m b_m\sigma_m\right) \\
&= \frac{1}{4}\sum_{k\ell mp} J_{k\ell}J_{mp}C_{k\ell mp} + \sum_{km} b_k b_m\delta_{km} = \frac{1}{2}\sum_{k\ell} J_{k\ell}^2 + \sum_k b_k^2.
\end{aligned} \tag{3.31}$$

Furthermore, the third term in the sum in Eq. (3.30) is given by

$$\langle\sigma_i H(\boldsymbol{\sigma})\rangle_{\beta=0} = -\frac{1}{2^n}\sum_{\{\boldsymbol{\sigma}\}}\sigma_i\left(\frac{1}{2}\sum_{k\ell} J_{k\ell}\sigma_k\sigma_\ell + \sum_k b_k\sigma_k\right) = -\sum_k b_k\delta_{ik} = -b_i. \tag{3.32}$$

Finally, we evaluate the first term in the sum in Eq. (3.30):

$$\begin{aligned}\left\langle \sigma_i \sigma_j H^2(\boldsymbol{\sigma}) \right\rangle_{\beta=0} &= \frac{1}{2^n} \sum_{\{\boldsymbol{\sigma}\}} \sigma_i \sigma_j \left(\frac{1}{2} \sum_{k\ell} J_{k\ell} \sigma_k \sigma_\ell + \sum_k b_k \sigma_k \right) \\ &\qquad \cdot \left(\frac{1}{2} \sum_{mp} J_{mp} \sigma_m \sigma_p + \sum_m b_m \sigma_m \right) \\ &= \frac{1}{4} \sum_{k\ell mp} J_{k\ell} J_{mp} C_{ijk\ell mp} + \sum_{km} b_k b_m C_{ijkm} \qquad (3.33) \\ &= \frac{1}{2} \delta_{ij} \sum_{k\ell} J_{k\ell}^2 + 2 \sum_k J_{ik} J_{kj} - 2\delta_{ij} \sum_k J_{jk}^2 + \delta_{ij} \sum_k b_k^2 + 2 b_i b_j - 2\delta_{ij} b_i^2.\end{aligned}$$

Combining all of the terms in Eq. (3.30) and canceling appropriately, we are left with

$$\begin{aligned}\frac{\partial^2}{\partial \beta^2} \left(\frac{1}{\beta} \chi_i \right)_{\beta=0} &= 2 \sum_j \left(\sum_k J_{ik} J_{kj} - \delta_{ij} \sum_k J_{jk}^2 - \delta_{ij} b_i^2 \right) \\ &= 2 \left(\sum_{j \neq i} (J^2)_{ij} - b_i^2 \right), \qquad (3.34)\end{aligned}$$

where we identify $\sum_{j \neq i} (J^2)_{ij}$ as the second degree of node i, i.e., the weight of paths of length two originating from i (not counting self-interactions).

All together, eqs. (3.27), (3.29), and (3.34) represent the zeroth-, first-, and second-order terms in the expansion of $\frac{1}{\beta}\chi_i$ in Eq. (3.21), respectively. Multiplying by β yields the desired high-temperature expansion of χ_i:

$$\chi_i = \beta + \beta^2 d_i + \beta^3 \left(\sum_{j \neq i} (J^2)_{ij} - b_i^2 \right) + \dots. \qquad (3.35)$$

3.8.2 *Low-temperature solution for general systems and general budgets*

We generalize the description of low-temperature optimal external fields provided in the main text to include general Ising systems (with degenerate low-energy states) and general external field budgets. We consider an Ising system described by the coupling matrix $J = J^T \in \mathbb{R}^{n \times n}$, an initial external field $\mathbf{b}^0 \in \mathbb{R}^n$, and temperature $T > 0$; and we consider a general external field budget $H > 0$. To avoid confusion between the budget H and the Hamiltonian, and to make the dependence on the additional external field explicit, in this section, for any feasible external field $\mathbf{h} \in \mathcal{F}_H$, we denote the Hamiltonian by

$$E_{\mathbf{h}}(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j - \sum_i (b_i^0 + h_i) \sigma_i. \qquad (3.36)$$

For every feasible external field $\mathbf{h} \in \mathcal{F}_H$, let $\Omega^0_{\mathbf{h}} = \arg\min_{\{\boldsymbol{\sigma}\}} E_{\mathbf{h}}(\boldsymbol{\sigma})$ denote the set of ground state configurations under the external field $\mathbf{b}^0 + \mathbf{h}$, and let $\Omega^0 = \cup_{\mathbf{h}\in\mathcal{F}_H}\Omega^0_{\mathbf{h}}$ denote the set of all possible ground states that can be induced by some $\mathbf{h} \in \mathcal{F}_H$.

For sufficiently low temperatures, the ground state of the system dominates the magnetization. Thus, any optimal external field will necessarily induce a ground state that has the largest magnetization among the configurations in Ω^0. So, let $\Omega^{0*} \subset \Omega^0$ denote the subset of possible ground states with the maximum magnetization and let $\mathcal{F}^*_H = \left\{\mathbf{h} \in \mathcal{F}_H \,:\, \Omega^0_{\mathbf{h}} \subset \Omega^{0*}\right\}$ denote the set of feasible external fields that induce ground states of maximum magnetization.

For sufficiently low temperatures, any optimal external field $\mathbf{h}^*$ is located in $\mathcal{F}^*_H$. To differentiate between the external fields in $\mathcal{F}^*_H$, we consider the possible first-excited states. Let $\Omega^1_{\mathbf{h}}$ denote the set of first excited states of the system under the external field $\mathbf{b}^0 + \mathbf{h}$, and let $\Omega^1 = \cup_{\mathbf{h}\in\mathcal{F}^*_H}\Omega^1_{\mathbf{h}}$ denote the set of possible first-excited states that can be induced by some $\mathbf{h} \in \mathcal{F}^*_H$. For every $\mathbf{h} \in \mathcal{F}^*_H$, we denote the energy gap between the ground and first-excited states under the external field $\mathbf{b}^0 + \mathbf{h}$ by:

$$\Delta E(\mathbf{h}) = \min_{\boldsymbol{\sigma}\in\Omega^1} \max_{\boldsymbol{\sigma}^0\in\Omega^{0*}} \left(E_{\mathbf{h}}(\boldsymbol{\sigma}) - E_{\mathbf{h}}(\boldsymbol{\sigma}^0)\right) \tag{3.37}$$

Letting $M_0 = \sum_i \sigma^0_i$, for any $\boldsymbol{\sigma}^0 \in \Omega^{0*}$, denote the magnetization of the optimal ground states, for any external field $\mathbf{h} \in \mathcal{F}^*_H$, the low-temperature magnetization takes the form:

$$M(\mathbf{b}^0 + \mathbf{h}) \approx \frac{\left|\Omega^0_{\mathbf{h}}\right| M_0 + \left(\sum_{\boldsymbol{\sigma}\in\Omega^1_{\mathbf{h}}} \sum_i \sigma_i\right) e^{-\beta\Delta E(\mathbf{h})}}{\left|\Omega^0_{\mathbf{h}}\right| + \left|\Omega^1_{\mathbf{h}}\right| e^{-\beta\Delta E(\mathbf{h})}} \approx M_0 + c(\mathbf{h}) e^{-\beta\Delta E(\mathbf{h})}, \tag{3.38}$$

where $c(\mathbf{h}) = \frac{1}{|\Omega^0_{\mathbf{h}}|}\left[\left(\sum_{\boldsymbol{\sigma}\in\Omega^1_{\mathbf{h}}} \sum_i \sigma_i\right) - \left|\Omega^1_{\mathbf{h}}\right| M_0\right]$ is a scalar that depends on $\mathbf{h}$. Thus, the low-temperature optimal external fields, i.e., the external fields which maximize the low-temperature magnetization, take the form:

$$\begin{aligned}
\mathbf{h}^* &= \arg\max_{\mathbf{h}\in\mathcal{F}^*_H} c(\mathbf{h}) e^{-\frac{1}{T}\Delta E(\mathbf{h})} \\
&\equiv \arg\max_{\mathbf{h}\in\mathcal{F}^*_H} -\operatorname{sign}\left[c(\mathbf{h})\right] \Delta E(\mathbf{h}) \\
&\equiv \arg\max_{\mathbf{h}\in\mathcal{F}^*_H} \min_{\boldsymbol{\sigma}\in\Omega^1} \max_{\boldsymbol{\sigma}^0\in\Omega^{0*}} -\operatorname{sign}\left[c(\mathbf{h})\right] \left(E_{\mathbf{h}}(\boldsymbol{\sigma}) - E_{\mathbf{h}}(\boldsymbol{\sigma}^0)\right)
\end{aligned} \tag{3.39}$$

Despite the complicated nature of Eq. (3.39), it reveals insight into the structure of low-T optimal external fields in general Ising systems with general external field budgets. Depending on the sign of $c(\mathbf{h})$, the optimal external field $\mathbf{h}^*$ will either maximize or minimize the energy gap. Since the first excited states in $\Omega^1_{\mathbf{h}}$ are likely ground states under other external fields (i.e., it is likely that $\Omega^1_{\mathbf{h}} \subset \Omega^0$ for $\mathbf{h} \in \mathcal{F}^*_H$), and since M_0 is the largest magnetization among the states in Ω^0, we should expect in most cases that $c(\mathbf{h}) < 0$ for all $\mathbf{h} \in \mathcal{F}^*_H$. In this case, the low-T optimal external

Algorithm 2: Projected gradient ascent

Input: An Ising system described by J, $\mathbf{b}^0$, and T; an external field budget H; and an accuracy parameter $\epsilon > 0$

Output: An external field $\mathbf{h}$ that approximates a local maximum of $M(\mathbf{b}^0 + \mathbf{h})$ in $\mathcal{F}_H$

$k = 0$; $\mathbf{h}^{(0)} \in \mathcal{F}_H$; $\alpha \in (0, \frac{1}{L})$;

repeat

  $\mathbf{h}^{(k+1)} = P_{\mathcal{F}_H}\left[\mathbf{h}^{(k)} + \alpha\chi(\mathbf{b}^0 + \mathbf{h}^{(k)})\right]$;

  $k{+}{+}$;

until $\left|M(\mathbf{b}^0 + \mathbf{h}^{(k+1)}) - M(\mathbf{b}^0 + \mathbf{h}^{(k)})\right| \leqslant \epsilon$;

$\mathbf{h} = \mathbf{h}^{(k+1)}$;

field $\mathbf{h}^*$ maximizes the energy gap $\Delta E(\mathbf{h})$, focusing H on nodes with opposite parity between the ground and first-excited states. Thus, once we restrict our attention to external fields that induce ground states with the maximum magnetization (i.e., once we restrict to $\mathbf{h} \in \mathcal{F}_H^*$), much of the intuition developed in the main text generalizes naturally to general Ising systems with general budgets.

3.8.3 *A projected gradient ascent algorithm*

We present a projected gradient ascent algorithm that uses the magnetic susceptibility to efficiently calculate local maxima of the magnetization for general Ising systems. The algorithm is initialized at a feasible external field $\mathbf{h}^{(0)} \in \mathcal{F}_H$ and steps along the direction of the gradient of the magnetization, which has a natural physical interpretation as the susceptibility $\chi(\mathbf{b}) = \nabla M(\mathbf{b})$. Projecting back onto $\mathcal{F}_H$ completes one iteration:

$$\mathbf{h}^{(k+1)} \leftarrow P_{\mathcal{F}_H}\left[\mathbf{h}^{(k)} + \alpha\chi(\mathbf{b}^0 + \mathbf{h}^{(k)})\right], \tag{3.40}$$

where $P_{\mathcal{F}_H}$ denotes the projection onto $\mathcal{F}_H$, which is well defined for norms $p \geqslant 1$, and $\alpha \geqslant 0$ is the step size. If the step size is chosen such that $\alpha \in (0, \frac{1}{L})$, where L is a Lipschitz constant of $M(\mathbf{b}^0 + \mathbf{h})$ (which is well-defined since M is smooth for finite systems), Algorithm 1 converges to an ϵ-approximation to a local maximum of $M(\mathbf{b}^0 + \mathbf{h})$ in $O(1/\epsilon)$ iterations (see (*657*)). Pseudocode for the algorithm is given in Algorithm 1.

We remark that, while Algorithm 1 is efficient in that it converges to a ϵ-approximate local maximum in $O(1/\epsilon)$ iterations, the total run-time is limited by the calculation of χ at each step. Since exact Ising calculations involve sums over $\{\pm 1\}^n$, which is exponential in the size of the system, an exact implementation of Algorithm 1 is limited to relatively small networks. To overcome this exponential dependence on system size, in the main text we use Monte Carlo simulations to approximate χ at each iteration.

Finally, we mention a special case in which the algorithm often converges to a global maximum of the magnetization; namely, when the couplings are ferromagnetic ($J \geqslant 0$)

and the initial external field is nonnegative ($\mathbf{b}^0 \geqslant 0$). We first note that the Fortuin-Kasteleyn-Ginibre inequality (222) states that, for any ferromagnetic system, regardless of external field, $\frac{\partial \langle \sigma_i \rangle}{\partial b_j} \geqslant 0$ for all $i, j \in N$, and hence $M(\mathbf{b}^0 + \mathbf{h})$ is non-decreasing in $\mathbf{h}$. This implies two important facts: (i) there exists a global maximum in $\mathcal{F}_H$ for which $\mathbf{h} \geqslant 0$, and (ii) if Algorithm 1 is initialized such that $\mathbf{h}^{(0)} \geqslant 0$, then we will have $\mathbf{h}^{(k)} \geqslant 0$ for all subsequent k. Secondly, the Griffiths-Hurst-Sherman inequality (279) states that, for any ferromagnetic system in a nonnegative external field, $\frac{\partial^2 \langle \sigma_i \rangle}{\partial b_j \partial b_k} \leqslant 0$ for all $i, j, k \in N$, and hence $M(\mathbf{b}^0 + \mathbf{h})$ is *entry-wise* concave in $\mathbf{h}$ for $\mathbf{h} \geqslant 0$. One would like to use this result to state that any local maximum of $M(\mathbf{b}^0 + \mathbf{h})$ for $\mathbf{h} \geqslant 0$ is a global maximum and, hence, our algorithm converges to a global maximum. However, such a statement requires that the Hessian $\frac{\partial^2 M}{\partial b_i \partial b_j}$ be *negative semidefinite*; i.e., all eigenvalues must be non-positive. While the theory doesn't quite guarantee convergence to a global maximum, in practice we find that the algorithm does often converge to a global maximum for ferromagnetic systems in non-negative external fields.

4

MAXIMIZING ACTIVITY IN ISING SYSTEMS VIA THE TAP APPROXIMATION

This chapter contains work from Lynn, Christopher W., and Daniel D. Lee. "Maximizing activity in Ising networks via the TAP approximation." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.

Abstract

A wide array of complex biological, social, and physical systems have recently been shown to be quantitatively described by Ising models, which lie at the intersection of statistical physics and machine learning. Here, we study the fundamental question of how to optimize the state of a networked Ising system given a budget of external influence. In the continuous setting where one can tune the influence applied to each node, we propose a series of approximate gradient ascent algorithms based on the Plefka expansion, which generalizes the naïve mean field and TAP approximations. In the discrete setting where one chooses a small set of influential nodes, the problem is equivalent to the famous influence maximization problem in social networks with an additional stochastic noise term. In this case, we provide sufficient conditions for when the objective is submodular, allowing a greedy algorithm to achieve an approximation ratio of $1 - 1/e$. Additionally, we compare the Ising-based algorithms with traditional influence maximization algorithms, demonstrating the practical importance of accurately modeling stochastic fluctuations in the system.

4.1 INTRODUCTION

The last 10 years have witnessed a dramatic increase in the use of maximum entropy models to describe a diverse range of real-world systems, including networks of neurons in the brain (*241*, *589*), flocks of birds in flight (*89*), and humans interacting in social networks (*237*, *418*), among an array of other social and biological applications (*357*, *402*, *457*, *523*). Broadly speaking, the maximum entropy principle allows scientists to formalize the hypothesis that large-scale patterns in complex systems emerge organically from an aggregation of simple fine-scale interactions between individual elements (*339*). Indeed, intelligence itself, either naturally-occurring in the human brain and groups of animals (*312*) or artificially constructed in learning algorithms and autonomous systems (*433*, *467*), is increasingly viewed as an emergent phenomenon

(*400*), the result of repeated underlying interactions between populations of smaller elements.

Given the wealth of real-world systems that are quantitatively described by maximum entropy models, it is of fundamental practical and scientific interest to understand how external influence affects the dynamics of these systems. Fortunately, all maximum entropy models are similar, if not formally equivalent, to the Ising model, which has roots in statistical physics (*113*) and has a rich history in machine learning as a model of neural networks (*160*). The state of an Ising system is described by the average activity of its nodes. For populations of neurons in the brain, an active node represents a spiking neuron while inactivity represents silence. In the context of humans in social networks, node activity could represent the sending of an email or the consumption of online entertainment, while inactivity represents moments in which an individual does not perform an action. By applying external influence to a particular node, one can shift the average activity of that node. Furthermore, this targeted influence also has indirect effects on the rest of the system, mediated by the underlying network of interactions. For example, if an individual is incentivized to send more emails, this shift in behavior induces responses from her neighbors in the social network, resulting in increased activity in the population as a whole.

As a first step toward understanding how to control such complex systems, we study the problem of maximizing the total activity of an Ising network given a budget of external influence. This so-called *Ising influence maximization* problem was originally proposed in the context of social networks (*415*), where it has a clear practical interpretation: If a telephone company or an online service wants to maximize user activity, how should it distribute its limited marketing resources among its customers? However, we emphasize that the broader goal—to develop a unifying control theory for understanding the effects of external influence in complex systems—could prove to have other important applications, from guiding healthy brain development (*257*) and intervening to alleviate diseased brain states (*256*) to anticipating trends in financial markets (*424*) and preventing viral epidemics (*513*).

We divide our investigation into two settings: (i) the continuous setting where one can tune the influence applied to each node, and (ii) the discrete setting in which one forces activation upon a small set of influential nodes. In the continuous setting, we propose a gradient ascent algorithm and give novel conditions for when the objective is concave. We then present a series of increasingly-accurate approximation algorithms based on an advanced approximation technique known as the Plefka expansion (*529*). The Plefka expansion generalizes the naïve mean field and TAP approximations, and, in theory, can be extended to arbitrary order (*717*).

In the discrete setting, it was recently shown that Ising influence maximization is closely related to the famous influence maximization problem in social networks (*365*) with the addition of a natural stochastic noise term (*416*). Here, we provide novel conditions for when the total activity of the system is submodular with respect to activated nodes. This result guarantees that a greedy algorithm achieves a $1 - 1/e$ approximation to the optimal choice of nodes. We compare our greedy algorithm

with traditional influence maximization techniques, demonstrating the importance of accurately accounting for stochastic noise.

Related work

Ising influence maximization was originally proposed in the context of human activity in social networks (*415*). However, a recent surge in the use of Ising models to describe other biological and physical systems significantly expands the problem's applicability (*631*).

Ising influence maximization was originally studied in the continuous setting under the naïve mean field approximation. Since the Plefka expansion generalizes the mean field approximation to increasing levels of accuracy, our work represents a principled improvement over existing techniques.

In the discrete setting, it was recently shown that Ising influence maximization is closely related to standard influence maximization (*416*), which was first studied in the context of viral marketing (*190*). Kempe et al. (*365*) proposed influence maximization as a discrete optimization problem and presented a greedy algorithm with approximation guarantees. Significant subsequent research has focused on developing efficient greedy and heuristic techniques (*139*, *140*, *396*). Here, we do not claim to provide improvements over these algorithms in the context of standard influence maximization. Instead, we focus on developing analogous techniques that are suitable for the Ising model.

4.2 ISING INFLUENCE MAXIMIZATION

In the study of complex systems, if we look through a sufficiently small window in time, the actions of individual elements appear binary—either human i sent an email ($\sigma_i = 1$) or she did not ($\sigma_i = -1$). The binary vector $\boldsymbol{\sigma} = \{\sigma_i\} \in \{\pm 1\}^n$ represents the activity of the entire system at a given moment in time, where n is the size of the system.

Many complex systems in the biological and social sciences have recently been shown to be quantitatively described by the Ising model from statistical physics. The Ising model is defined by the Boltzmann distribution over activity vectors:

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\frac{1}{2}\sum_{i \neq j} J_{ij}\sigma_i\sigma_j + \sum_i b_i\sigma_i\right), \tag{4.1}$$

where Z is a normalization constant. The parameters $J = \{J_{ij}\}$ define the network of interactions between elements and the parameters $\mathbf{b} = \{b_i\}$ represent individual biases, which can be altered by application of targeted external influence. For example, if J defines the network of interactions in a population of email users, then $\mathbf{b}$ represents the intrinsic tendencies of users to send emails, which can be shifted by incentivizing or disincentivizing email use.

Problem statement

The total average activity of a network with bias $\mathbf{b}$ is denoted $M(\mathbf{b}) = \sum_i \langle \sigma_i \rangle$, where $\langle \cdot \rangle$ denotes an average over the Boltzmann distribution (6.5). In what follows, we assume that the interactions J and initial bias $\mathbf{b}^0$ are known. We note that this assumption is not restrictive since there exist an array of advanced techniques in machine learning (3) and statistical mechanics (33) for learning Ising parameters directly from observations of a system.

We study the problem of maximizing the total activity M with respect to an additional external influence $\mathbf{h}$, subject to the budget constraint $|\mathbf{h}|_p = (\sum_i |h_i|^p)^{1/p} \leqslant H$, where H is the budget of external influence.

Problem 1 (Ising influence maximization). Given an Ising system defined by J and $\mathbf{b}^0$, and a budget H, find an optimal external influence $\mathbf{h}^*$ satisfying

$$\mathbf{h}^* = \underset{|\mathbf{h}|_p \leqslant H}{\arg\max}\, M(\mathbf{b}^0 + \mathbf{h}). \tag{4.2}$$

We point out that the norm p plays an important role. If $p = 1, 2, 3, \ldots$, then one is allowed to tune the influence on each node continuously. On the other hand, if $p = 0$, then $|\mathbf{h}|_0$ counts the number of non-zero elements in $\mathbf{h}$. In this case, one chooses a subset of $\lfloor H \rfloor$ nodes $\{i\}$ to activate with probability one by sending $\{h_i\} \to \infty$.

4.3 THE PLEFKA EXPANSION

Since the Ising model has remained unsolved for all but a select number of special cases, tremendous interdisciplinary effort has focused on developing tractable approximation techniques. Here, we present an advanced approximation method known as the Plefka expansion (717). The Plefka expansion is not an approximation itself, but is rather a principled method for deriving a series of increasingly accurate approximations, the first two orders of which are the naïve mean-field (MF) and Thouless-Anderson-Palmer (TAP) approximations. In subsequent sections, we will use the Plefka expansion to approximately solve the Ising influence maximization problem in (4.2).

Calculations in the Ising model, such as the average activity $\langle \sigma_i \rangle$, generally require summing over all 2^n binary activity vectors. To get around this exponential dependence on system size, the Plefka expansion provides a series of approximations based on the limit of weak interactions $|J_{ij}| \ll 1$. Each order α of the expansion generates a set of self-consistency equations $m_i = f_i^{(\alpha)}(\mathbf{m})$, where m_i approximates the average activity $\langle \sigma_i \rangle$. Thus, for any order α of the Plefka expansion, the intractable problem of computing the averages $\langle \sigma_i \rangle$ is replaced by the manageable task of computing solutions to the corresponding self-consistency equations $\mathbf{m} = \mathbf{f}^{(\alpha)}(\mathbf{m})$. We point the interested reader to Sec. 4.7.1 for a detailed derivation of the Plefka expansion.

For a system with interactions J and bias $\mathbf{b}$, the first order in the expansion yields the naïve mean field approximation, summarized by the self-consistency equations

$$m_i = \tanh\left[b_i + \sum_j J_{ij} m_j\right] \triangleq f_i^{\mathrm{MF}}(\mathbf{m}). \tag{4.3}$$

The second order in the Plefka expansion yields the TAP approximation,

$$m_i = \tanh\left[b_i + \sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1 - m_j^2)\right] \triangleq f_i^{\mathrm{TAP}}(\mathbf{m}). \tag{4.4}$$

Higher-order approximations can be achieved by systematically including higher orders of J in the argument of $\tanh[\cdot]$. In Sec. 4.7.2, we present a derivation of the third-order approximation, denoted TAP3.

The standard approach for computing solutions to the self-consistency equations $\mathbf{m} = \mathbf{f}^{(\alpha)}(\mathbf{m})$ is to iteratively apply $\mathbf{f}^{(\alpha)}$ until convergence is reached:

$$\mathbf{m} \leftarrow (1-\gamma)\mathbf{m} + \gamma \mathbf{f}^{(\alpha)}(\mathbf{m}), \tag{4.5}$$

where $\gamma \in [0, 1]$ is the step size. The convergence of this procedure was rigorously examined in (*95*). In practice, we find that $\gamma \sim 0.01$ yields rapid convergence for most systems up to the third-order approximation.

4.4 THE CONTINUOUS SETTING

In this section, we study Ising influence maximization under a budget constraint $|\mathbf{h}|_p \leqslant H$, where $p = 1, 2, 3, \ldots$, yielding a continuous optimization problem where one can tune the external influence on each element in the system. We first present an exact gradient ascent algorithm and comment on its theoretical guarantees. We then demonstrate how the Plefka expansion can be used to approximate the gradient, yielding a series of increasingly accurate approximation algorithms.

4.4.1 *Projected gradient ascent*

We aim to maximize the total activity $M(\mathbf{b}^0 + \mathbf{h}) = \sum_i \langle \sigma_i \rangle$ with respect to the external influence $\mathbf{h}$. Thus, a crucially important concept is the response function

$$\chi_{ij} = \frac{\partial \langle \sigma_i \rangle}{\partial h_j} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle, \tag{4.6}$$

which quantifies the change in the activity of node i due to a shift in the influence on node j. The gradient of M with respect to $\mathbf{h}$ can be succinctly written $\nabla_{\mathbf{h}} M = \chi^{\mathsf{T}} \mathbf{1}$, where $\mathbf{1}$ is the n-vector of ones.

Algorithm 3: Projected Gradient Ascent (PGA)

Input: Ising system defined by J and $\mathbf{b}^0$, budget H, norm p, and error ϵ;
Output: External influence $\mathbf{h}$;
Initialize: Choose $|\mathbf{h}^{(0)}|_p \leqslant H$, $k \leftarrow 0$;
while $\left|M(\mathbf{b}^0 + \mathbf{h}^{(k)}) - M(\mathbf{b}^0 + \mathbf{h}^{(k-1)})\right| > \epsilon$ **do**
  Choose step size η_k; $\mathbf{h}^{(k+1)} \leftarrow \pi_{|\mathbf{h}|_p \leqslant H}\left[\mathbf{h}^{(k)} + \eta_k \chi(\mathbf{b}^0 + \mathbf{h}^{(k)})^T \mathbf{1}\right]$;
  k++;
end
$\mathbf{h} \leftarrow \mathbf{h}^{(k)}$;

In Algorithm 1 we present a projected gradient ascent algorithm PGA. Starting at a feasible choice for the external influence $\mathbf{h}^{(0)}$, PGA steps along the gradient $\chi^T\mathbf{1}$ and then projects back down to the space of feasible solutions $|\mathbf{h}|_p \leqslant H$. We note that for $p = 1, 2, 3, \ldots$, the space of feasible solutions is convex, and hence the projection $\pi_{|\mathbf{h}|_p \leqslant H}$ is well-defined and can be performed efficiently (*196*).

4.4.2 *Conditions for optimality*

The algorithm PGA efficiently converges to an ϵ-approximation of a local maximum of M in $O(1/\epsilon)$ iterations (*471*). However, this local maximum could be arbitrarily far from the globally optimal solution. Here, we present a novel sufficient condition for when PGA is guaranteed to converge to a global maximum of M, subject to the proof of a long-standing conjecture.

Conjecture 2 (*647*)**.** Given an Ising system with non-negative interactions $J \geqslant 0$ and non-negative biases $\mathbf{b} \geqslant 0$, the average activity of each node $\langle\sigma_i\rangle$ is a concave function of the biases $\mathbf{b}$.

Theorem 3. If Conjecture 2 holds, then for any Ising system with non-negative interactions $J \geqslant 0$ and non-negative initial biases $\mathbf{b}^0 \geqslant 0$, PGA converges to a global maximum of the total activity M.

Proof. For Ising systems with positive couplings $J \geqslant 0$, the response function is non-negative $\{\chi_{ij}\} \geqslant 0$ (*280*). This implies two things: (i) at least one global maximum of $M(\mathbf{b})$ occurs in the non-negative orthant of $\mathbf{b}$, and (ii) if $\mathbf{b}^0 \geqslant 0$, then $\mathbf{b}^0 + \mathbf{h}^{(k)}$ will be non-negative at every iteration k of PGA. If Conjecture 2 holds, then every local maximum in the non-negative orthant is a global maximum. Thus, PGA converges to a global maximum. □

We remark that Sylvester (*647*) provides extensive experimental justification for Conjecture 2, and even proves Conjecture 2 in a number of limited cases. Additionally,

we manually verified the veracity of Conjecture 2 in each of the experiments presented below. We also note that the sufficient conditions are plausible for many real-world scenarios. Positive interactions $J \geqslant 0$ imply that an action from one node will tend to induce an action from another node, a phenomenon that has been experimentally verified in small neuronal (*589*) and social (*418*) networks. The more stringent condition is that $\mathbf{b}^0 \geqslant$, implying that each element in the network prefers activity over inactivity.

4.4.3 *Approximating the gradient via the Plefka expansion*

Since PGA requires calculating the response function χ at each iteration, an exact implementation scales exponentially with the size of the system. Here we show that the Plefka expansion can be used to approximate χ, yielding a series of efficient and increasingly-accurate gradient ascent algorithms.

Given a self-consistent approximation of the form $\mathbf{m} = \mathbf{f}^{(\alpha)}(\mathbf{m})$, where α denotes the order of the Plefka approximation, the response function is approximated by

$$\tilde{\chi}^{(\alpha)}_{ij} = \frac{\partial f^{(\alpha)}_i}{\partial h_j} + \sum_k \frac{\partial f^{(\alpha)}_i}{\partial m_k} \tilde{\chi}^{(\alpha)}_{kj}. \tag{4.7}$$

For all orders α of the Plefka expansion, we point out that $\partial f^{(\alpha)}_i / \partial h_j = (1 - m_i^2)\delta_{ij}$. Thus, defining $A_{ij} \triangleq (1 - m_i^2)\delta_{ij}$, and denoting the Jacobian of $\mathbf{f}^{(\alpha)}$ by $Df^{(\alpha)}_{ij} \triangleq \partial f^{(\alpha)}_i / \partial m_j$, the response function takes the particularly simple form

$$\tilde{\chi}^{(\alpha)} = (I - Df^{(\alpha)})^{-1} A, \tag{4.8}$$

where I is the identity matrix.

Thus, to approximate the gradient $\nabla_{\mathbf{h}} M \approx \tilde{\chi}^T \mathbf{1}$, one simply needs to calculate the Jacobian of $\mathbf{f}^{(\alpha)}$. Under the mean field approximation, the Jacobian takes the form $Df^{MF} = AJ$; and under the TAP approximation, we have

$$Df^{TAP}_{ij} = (1 - m_i^2)\Big[J_{ij} + 2J_{ij} m_i m_j - \delta_{ij} \sum_k J_{ik}(1 - m_k^2)\Big]. \tag{4.9}$$

We point the reader to Sec. 4.7.2 for a derivation of the third-order Jacobian.

4.4.4 *Experimental evaluation*

In Fig. 4.1, we compare various orders of the Plefka approximation across a range of networks for the norm $p = 1$. We also compare with the uniform influence $\mathbf{h} = H/n\mathbf{1}$ as a baseline. In Fig. 4.1a, the network is small enough that we can calculate the exact optimal solution $\mathbf{h}^*$, while for Fig. 4.1b-d, we approximate $\mathbf{h}^*$ by running costly Monte Carlo simulations to estimate the gradient at each iteration of PGA. Similarly, we

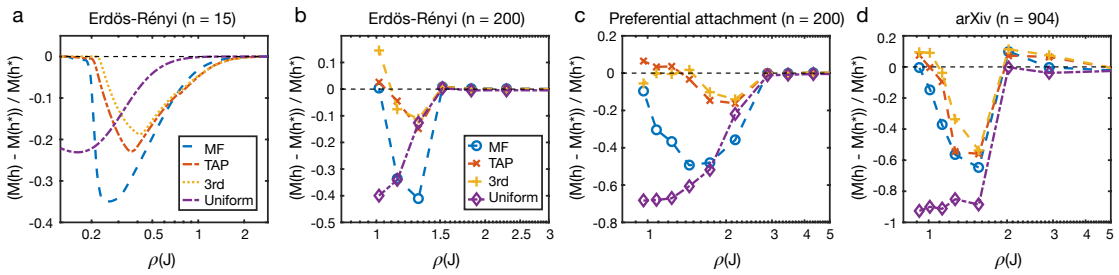


Figure 4.1: Performance of PGA for various orders of the Plefka expansion. (*a*) An Erdös-Rényi network with $n = 15$ nodes and budget $H = 1$. The total activity is calculated exactly using the Boltzmann distribution. (*b*) An Erdös-Rényi network with $n = 200$ nodes and budget $H = 10$. (*c*) A preferential attachment network with $n = 200$ nodes and budget $H = 10$. (*d*) A collaboration network of $n = 904$ physicists on the arXiv and budget $H = 20$. The total activities in (b-d) are estimated using Monte Carlo simulations. The benchmarks are PGA with the exact gradient for (*a*) and the gradient estimated using Monte Carlo simulations in (*b-d*).

calculate the total activity M exactly using the Boltzmann distribution in Fig. 4.1a, while in Fig. 4.1b-d, we estimate M using Monte Carlo simulations.

For each network, we assume that the interactions are symmetric $J = J^T$ with uniform weights and that the initial bias is zero $\mathbf{b}^0 = 0$. We then study the performance of the various algorithms across a range of interaction strengths, summarized by the spectral radius $\rho(J)$. For $\rho(J) \ll 1$, the network is dominated by randomness and all influence strategies have little affect on the total activity. On the other hand, for $\rho(J) \gg 1$, the elements interact strongly and any positive influence induces the entire network to become active. Thus, the interesting regime is the "critical" region near $\rho(J) \approx 1$.

The striking success of the Plefka expansion is summarized by the fact that TAP and TAP3 consistently provide dramatic improvements over the naïve mean field algorithm studied in (*415*). Indeed, TAP and TAP3 consistently perform within 20% of optimal (except for the arXiv network) and sometimes even outperform the Monte Carlo algorithm benchmark in Fig. 4.1b-d. Furthermore, while the Monte Carlo algorithm takes ~ 10 minutes to complete in a network of size 200, PGA with the TAP and TAP3 approximations converges within ~ 5 seconds.

4.5 DISCRETE SETTING

We now consider the discrete setting corresponding to a budget constraint of the form $|\mathbf{h}|_0 \leqslant H$. In this setting, one is allowed to apply infinite external influence to a set of $\lfloor H \rfloor$ nodes in the system, activating them with probability one; that is, one chooses a set of nodes $V = \{i\}$ for which we impose $\langle \sigma_i \rangle = 1$ by taking $h_i \to +\infty$. We begin by presenting a greedy algorithm that selects the single node at each iteration that yields the largest increase in the total activity M. We then provide novel conditions for when M is submodular in the selected nodes, which guarantees that our greedy algorithm is within $1 - 1/e$ of optimal. Finally, we comment on the relationship between (discrete) Ising influence maximization and traditional influence maximization in

Algorithm 4: Greedy algorithm for choosing top H influential nodes in an Ising network (GI)

Input: Ising system defined by J and $\mathbf{b}^0$, budget H;
Output: Set of H influential nodes V;
Initialize: $V^{(1)} \leftarrow \{\}$;
for $k = 1, \ldots, H$ **do**
  for *all nodes* $i \in \{1, \ldots, n\}/V^{(k)}$ **do**
    Calculate total activity $\mathcal{M}(V^{(k)} \cup \{i\})$;
  end
  Choose node $i^* = \arg\max_i \mathcal{M}(V^{(k)} \cup \{i\})$;
  Add i^* to influential set $V^{(k+1)} \leftarrow V^{(k)} \cup \{i^*\}$;
end
$V \leftarrow V^{(k+1)}$;

viral marketing, and we present experiments comparing our greedy algorithm with traditional techniques.

4.5.1 *A greedy algorithm*

We aim to maximize the total activity M with respect to a set V of activated nodes of size $|V| = H$ (assuming H is integer). To eliminate confusion, we denote by $\mathcal{M}(V)$ the total activity of the system after activating the nodes in V, assuming that the couplings J and initial bias $\mathbf{b}^0$ are already known.

Since there are $\binom{n}{H} \sim n^H$ possible choices for V, an exhaustive search for the optimal set is generally infeasible. On the other hand, we can simplify our search by looking at one node at a time and iteratively adding to V the single node that increases $\mathcal{M}$ the most. This approximate greedy approach was made famous in traditional influence maximization in the context of viral marketing by Kempe et al. (*365*). In Algorithm 4 we propose an analogous algorithm for computing the top H influential nodes in an Ising system.

4.5.2 *Theoretical guarantee*

The greedy algorithm GI efficiently chooses an approximate set V of influential nodes in $O(nH)$ iterations. However, V could be arbitrarily far from the globally optimal set of nodes. Here, we present novel conditions for when $\mathcal{M}$ is monotonic and submodular in V, and, hence, GI is guaranteed to compute a $1 - 1/e$ approximation of the optimal set of nodes. The proof is based on the famous GHS inequality from statistical physics.

Theorem 4 (*279*)**.** Given an Ising system with non-negative interactions $J \geqslant 0$ and non-negative biases $\mathbf{b} \geqslant 0$, we have $\frac{\partial^2 \langle \sigma_i \rangle}{\partial b_j \partial b_k} \leqslant 0$ for all elements i, j, k.

We note that the GHS inequality guarantees a limited type of concavity of $\langle \sigma_i \rangle$ in the direction of positive bias $\mathbf{b}$. While this was not enough to prove that PGA is optimal in the continuous setting, it is strong enough to guarantee that $\mathcal{M}$ is submodular in the discrete setting.

Theorem 5. For an Ising system with non-negative interactions $J \geqslant 0$ and non-negative initial biases $\mathbf{b}^0 \geqslant 0$, the total activity $\mathcal{M}$ is monotonic and submodular in the activated nodes V.

Proof. Monotonicity is guaranteed for any system with non-negative interactions in (*280*). To prove submodularity, we first introduce the notation $h_i^V \in \{0, 1\}$ if i is or is not in V. Then we note that $\mathcal{M}(V) \equiv \lim_{c\to\infty} M(\mathbf{b}^0 + c\mathbf{h}^V)$. Since M is non-negative and concave in the direction of positive bias for $J \geqslant 0$ and $\mathbf{b}^0 \geqslant 0$, M it is subadditive. Thus, for any set V of nodes and any two nodes $i, j \notin V$, we have

$$\begin{aligned} &M(\mathbf{b}^0 + c(\mathbf{h}^V + \mathbf{h}^{\{i\}})) + M(\mathbf{b}^0 + c(\mathbf{h}^V + \mathbf{h}^{\{j\}})) \\ &\geqslant M(\mathbf{b}^0 + c(\mathbf{h}^V + \mathbf{h}^{\{i\}} + \mathbf{h}^{\{j\}})) + M(\mathbf{b}^0 + c\mathbf{h}^V). \end{aligned} \tag{4.10}$$

Taking $c \to \infty$, we find that

$$\mathcal{M}(V \cup \{i\}) + \mathcal{M}(V \cup \{j\}) \geqslant \mathcal{M}(V \cup \{i, j\}) + \mathcal{M}(V), \tag{4.11}$$

which is the formal definition for submodularity. □

4.5.3 *Relationship between the linear threshold and Ising models*

It was recently established that the discrete version of the Ising influence maximization problem is closely related to traditional influence maximization in social networks (*416*). In traditional influence maximization, one aims to maximize the spread of activations under a viral model, such as linear threshold (LT) or independent cascade. For example, the LT model is defined by the deterministic dynamics

$$\sigma_i^{(t+1)} \leftarrow \operatorname{sign}\Big[\sum_j J_{ij}\sigma_j^{(t)} + b_i\Big]. \tag{4.12}$$

Typically, one considers activation variables $\sigma_i' \in \{0, 1\}$ instead of $\sigma_i = \pm 1$, which can be accomplished by a simple change of parameters $J_{ij}' \leftarrow 2J_{ij}$ and $b_i' \leftarrow b_i - \sum_j J_{ij}$. The negative bias $\theta_i = -b_i'$ is referred to as the threshold of i, representing the amount of influence i must receive from her neighbors to become active.

We include a stochastic influence ϵ for each node at every iteration t of the LT dynamics, representing natural fluctuations in the influence on each node over time. If ϵ is drawn from a logistic distribution, then these stochastic dynamics are equivalent to Glauber Monte Carlo dynamics (*476*)

$$P\left(\sigma_i^{(t+1)} \middle| \boldsymbol{\sigma}^{(t)}\right) = \left(1 + e^{-\frac{2}{T}\left(\sum_j J_{ij}\sigma_j^{(t)} + b_i\right)}\right)^{-1}, \tag{4.13}$$

where T parameterizes the variance of ϵ. Allowing the system time to equilibrate, the statistics of the Glauber dynamics follow the Boltzmann distribution (noting that we have taken $T = 1$ in Eq. (6.5)), simulating an Ising model. Furthermore, it is clear to see that we recover the deterministic LT dynamics in the limit of zero fluctuations $T \to 0$. Thus, the Ising model represents a natural generalization of the LT model to settings with stochastic fluctuations in the influence. We emphasize that, because maximum entropy models have demonstrated tremendous ability in quantitatively describing a wide range of real-world systems (*89*, *241*, *357*, *523*, *589*, *631*), understanding how these systems react to external influence is a significant endeavor in and of itself, and this goal should fundamentally be viewed as running adjacent to, as opposed to in conflict with, the existing viral influence maximization literature.

Finally, we note that most applications of the LT model to influence maximization impose the constraint $\sum_j J'_{ij} \leqslant 1$ ($\sum_j J_{ij} \leqslant 1/2$ in Ising notation) and assume that the bias b'_i is drawn uniformly from $[0, 1]$ ($b_i \sim \mathcal{U}[-1/2, 1/2]$) at the beginning of each simulation. Randomly selecting b_i at the beginning of each simulation is meant to represent uncertainty in the nodes' biases, which is fundamentally distinct from randomizing b_i at *each iteration* to represent natural xs in the biases over time. Indeed, in the following experiments we include both sources of randomness, making our model equivalent to the so-called random-field Ising model, which shows similar behavior to a spin glass (*468*).

4.5.4 *Experimental evaluation*

We experimentally evaluate the performance of our greedy algorithm under various orders of the Plefka expansion. We also provide the first comparison between Ising influence algorithms and the traditional greedy influence maximization algorithm in (*365*). As is usually assumed in traditional influence maximization, we scale the interactions such that $\sum_j J_{ij} \leqslant 1/2$. Furthermore, we draw the initial bias on each node from a uniform distribution $b_i \sim \mathcal{U}[-1/2, 1/2]$ and average over many such draws.

To fairly compare the Ising and linear threshold algorithms, we divide the experiments into two classes: the first is evaluated with respect to the total activity $\mathcal{M}$ under the Ising model, while the second class of experiments evaluates the spread S resulting from each choice of nodes under the linear threshold model. We denote the greedy algorithm in (*365*) by IM. We also compare with the heuristic strategy of choosing the top H nodes with the highest degrees in the network.

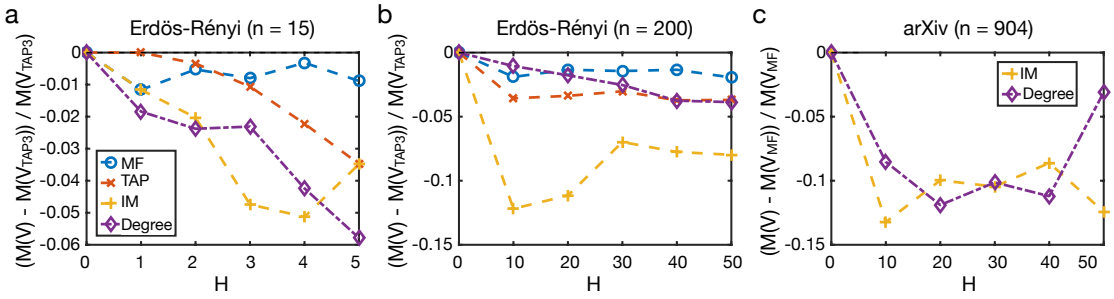


Figure 4.2: Comparison of the total Ising activity for greedy algorithms using various orders of the Plefka expansion. For each network, we ensure $\sum_j J_{ij} \leqslant 1/2$ and we average over many draws of the initial bias $\{b_i^0\} \sim \mathcal{U}[-1/2, 1/2]$. (*a*) An Erdös-Rényi network with $n = 15$ nodes. The total activity is calculated exactly using the Boltzmann distribution. (*b*) An Erdös-Rényi network with $n = 200$ nodes. (*c*) A collaboration network of $n = 904$ physicists on the arXiv. The total activities in (*b-c*) are estimated using Monte Carlo simulations. In (*a-b*) the benchmark is TAP3, while for (*c*) the benchmark is MF.

4.5.4.1 *Comparison under the Ising model*

We first compare the different greedy algorithms under the Ising model. In Fig. 4.2a,b, we use the Ising greedy algorithm GI with the third-order approximation TAP3 as the benchmark. In both Erdös-Rényi networks, we find that GI with TAP3 slightly outperforms TAP and MF, while all three Ising-based algorithms significantly out-perform the linear-threshold-based algorithm IM and the degree heuristic. Since TAP3, TAP, and MF all perform within 5% of one another, in Fig. 4.2c we use GI with the MF approximation as the benchmark. In the arXiv network, we find that GI with MF significantly outperforms both LT and the degree heuristic. These results demonstrate the practical importance of accurately modeling the stochastic noise in the system. We point out that the total activity $\mathcal{M}$ is calculated exactly in Fig. 4.2a using the Boltzmann distribution and estimated in Fig. 4.2b,c using Monte Carlo simulations.

4.5.4.2 *Comparison under the linear threshold model*

We now compare the algorithms under the linear threshold model. In all of Fig. 4.3a-c, we use the LT greedy algorithm IM as the benchmark and exactly compute the spread of influence under the LT model. Surprisingly, in both of the Erdös-Rényi networks in Fig. 4.3a,b, all of the Ising-based algorithms and the degree heuristic out-perform IM. In particular, the third-order approximation TAP3 achieves close to the largest spread in both networks. In the arXiv network in Fig. 4.3c, the Ising-based algorithm continues to slightly out-perform IM, while IM out-performs the degree heuristic.

The success of the Ising-based algorithms is surprising, since they are all attempting to maximize a fundamentally different objective from influence spread under LT. We suspect that the strong performance might be the result of the Ising model performing a type of simulated annealing, similar to recent techniques proposed in (*342*); however, an investigation of this hypothesis is beyond the scope of the current paper.

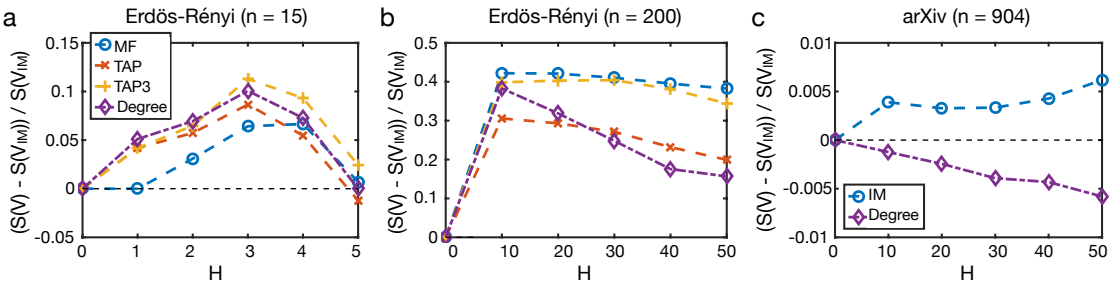


Figure 4.3: Comparison of the spread of influence under the linear threshold model for different greedy algorithms. For each network, we ensure $\sum_j J_{ij} \leqslant 1/2$ and we average over many draws of the initial bias $\{b_i^0\} \sim \mathcal{U}[-1/2, 1/2]$. (*a*) An Erdös-Rényi network with $n = 15$ nodes. (*b*) An Erdös-Rényi network with $n = 200$ nodes. (*c*) A collaboration network of $n = 904$ physicists on the arXiv. The benchmark in all panels is IM.

4.6 CONCLUSIONS

Maximum entropy models such as the Ising model have recently been used to quantitatively describe a plethora of biological, physical, and social systems. Given the success of the Ising model in capturing real-world systems, including populations of neurons in the brain and networks of interacting humans, understanding how to control and optimize the large-scale behavior of complex systems is of fundamental practical and scientific interest, with applications from guiding healthy brain development (*257*) and intervening to alleviate diseased brain states (*256*) to anticipating trends in financial markets (*424*) and preventing viral epidemics (*513*).

Here, we study the problem of maximizing the total activity of an Ising network given a budget of external influence. In the continuous setting where one can tune the influence on each node, we present a series of increasingly-accurate gradient ascent algorithms based on an approximation technique known as the Plefka expansion. In the discrete setting where one chooses a set of influential nodes, we propose a greedy algorithm and present novel conditions for when the objective is submodular.

Future work

Given the novelty of this problem and the recent surge in the use of the Ising model, there are many promising directions to pursue. One direction is to consider a more general control problem in which the controller aims to shift the Ising network into a desired state instead of simply maximizing the activity of all nodes.

Another direction is to consider data-based optimization, wherein the optimizer is only aware of the past activity of the system (*274*). Since the Ising model is mathematically equivalent to a Boltzmann machine (*160*), one could adapt state-of-the-art methods from machine learning to approach this problem.

Finally, given the experimental success of the Ising-based greedy algorithms under the linear threshold model, an obvious extension of the current work should look into

possible explanations. We suspect that a closer comparison with simulated annealing techniques in (*342*) might provide the answer.

4.7 SUPPLEMENTARY MATERIAL

4.7.1 *The Plefka expansion*

Many complex systems in the biological and social sciences have recently been shown to be quantitatively described by the Ising model from statistical physics (*237*, *241*, *418*, *589*). The Ising model is defined by the Boltzmann distribution:

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(-H(\boldsymbol{\sigma})\right), \tag{4.14}$$

$$H(\boldsymbol{\sigma}) = -\frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j - \sum_i b_i \sigma_i, \text{and} \tag{4.15}$$

$$Z = \sum_{\boldsymbol{\sigma} \in \{\pm 1\}^n} e^{-\beta H(\boldsymbol{\sigma})}, \tag{4.16}$$

where $H(\boldsymbol{\sigma})$ is the energy function, or Hamiltonian, that defines the system, and Z is a normalization constant. The parameters $J = \{J_{ij}\}$ define the network of interactions between elements and the parameters $\mathbf{b} = \{b_i\}$ represent individual biases, which can be altered by application of targeted external influence.

Since the Ising model has remained unsolved for all but a select number of special cases, tremendous interdisciplinary effort has focused on developing tractable approximation techniques. Here, we present an advanced approximation method known as the Plefka expansion (*717*). The Plefka expansion is not an approximation itself, but is rather a principled method for deriving a series of increasingly accurate approximations, the first two orders of which are the naïve mean-field (MF) and Thouless-Anderson-Palmer (TAP) approximations.

The Plefka expansion is a small-interaction expansion derived by Georges and Yedidia in (*244*), extending the work of Thouless, Anderson, Palmer, and Plefka (*529*, *664*). Throughout this section, we assume that the interactions are symmetric (i.e., $J = J^T$) to ease the presentation; however, we note that the Plefka expansion easily generalizes to include asymmetric interactions as well (*415*, *726*). Furthermore, because we are expanding in the limit $J \ll 1$, it is useful to take $J \to \beta J$, where β parameterizes the strength of interactions in the system.

Given an Ising system defined by J, $\mathbf{b}$, and β, we consider the free energy $G = -\ln(Z)/\beta$, which can be thought of as a mathematical tool whose derivative with respect to the external field b_i defines the average activity of node i; i.e., $\langle \sigma_i \rangle = -\partial G / \partial b_i$. To derive the Plefka expansion, we consider the approximate free energy:

$$\beta \tilde{G} = -\ln \sum_{\{\boldsymbol{\sigma}\}} \exp\Big(-\beta H(\boldsymbol{\sigma}) + \sum_i \lambda_i (\sigma_i - m_i) \Big), \tag{4.17}$$

where λ_i are auxiliary fields that impose the constraints $m_i = \langle \sigma_i \rangle$ and are eventually set to zero to recover the true free energy. Using (4.17), we can expand the free energy

around the small-interaction limit $\beta = 0$. Carrying out this expansion to third-order in β, we obtain (*717*):

$$\begin{aligned}\beta G = &\sum_i \frac{1+m_i}{2}\ln\left(\frac{1+m_i}{2}\right) + \frac{1-m_i}{2}\ln\left(\frac{1-m_i}{2}\right) \\ &- \beta\Big(\frac{1}{2}\sum_{ij} J_{ij} m_i m_j + \sum_i b_i m_i\Big) \\ &- \frac{\beta^2}{4}\sum_{ij} J_{ij}^2 (1-m_i^2)(1-m_j^2) \\ &- \frac{\beta^3}{3}\sum_{ij} J_{ij}^3 m_i (1-m_i^2) m_j (1-m_j^2) \\ &- \frac{\beta^3}{6}\sum_{ijk} J_{ij} J_{jk} J_{ki} (1-m_i^2)(1-m_j^2)(1-m_k^2) + \dots . \end{aligned} \tag{4.18}$$

The zeroth-order term in the expansion corresponds to the mean-field entropy and the first-order term is the mean-field energy. Thus, up to first-order in β, we recover the MF approximation. The second-order term is known as the Onsager reaction term, and its inclusion yields the TAP approximation (*664*). Higher-order terms are systematic corrections first presented in (*244*) and, in principle, can be carried out to arbitrary order.

The quantities m_i approximate the average activities $\langle \sigma_i \rangle$ and are defined by the stationary conditions $\partial \tilde{G}/\partial m_i = 0$. Thus, differentiating (4.18) with respect to m_i, and only keeping terms to second-order in β, we arrive at the TAP self-consistency equations for the magnetizations $\mathbf{m}$:

$$m_i = \tanh\Big[b_i + \sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1-m_j^2)\Big] \triangleq f_i^{\mathrm{TAP}}(\mathbf{m}). \tag{4.19}$$

Thus, for any order α of the Plefka expansion, the intractable problem of computing exact averages over the Glauber dynamics is replaced by the manageable task of computing a fixed point of the corresponding self-consistency map $\mathbf{m} = \mathbf{f}^{(\alpha)}(\mathbf{m})$.

4.7.2 *The third-order TAP approximation*

We now derive the self-consistency equations and response function for the third-order approximation in the Plefka expansion. To third-order in β, we arrive at the self-consistency equations:

$$m_i = \tanh\Bigg[b_i + \sum_j J_{ij} m_j - m_i \sum_j J_{ij}^2 (1-m_j^2) + \frac{2}{3}(1-3m_i^2) \sum_j J_{ij}^3 m_j (1-m_j^2) \\ - m_i \sum_{jk} J_{ij} J_{jk} J_{ki} (1-m_j^2)(1-m_k^2) \Bigg] \triangleq f_i^{\mathrm{TAP}_3}(\mathbf{m}). \tag{4.20}$$

For algorithmic purposes, we are also interested in the response function $\tilde{\chi}_{ij}^{(\alpha)} = \frac{\partial m_i}{\partial b_j}$, which, for any order α of the Plefka expansion, takes the form:

$$\tilde{\chi}^{(\alpha)} = (I - Df^{(\alpha)})^{-1} A, \tag{4.21}$$

where I is the identity matrix, $Df_{ij}^{(\alpha)} = \frac{\partial f_i^{(\alpha)}}{\partial m_j}$ is the Jacobian of the mean-field map, and $A_{ij} = (1-m_i^2)\delta_{ij}$ is a diagonal matrix. Thus, the response function for each extended mean-field approximation is defined by the Jacobian of the corresponding mean-field map. Up to third-order in β, Df^{TAP_3} takes the form:

$$Df_{ij}^{\mathrm{TAP}_3} = (1-m_i^2)\Bigg[J_{ij} + 2J_{ij}^2 m_i m_j - \delta_{ij} \sum_k J_{ik}^2 (1-m_k^2) \\ - 4m_i \delta_{ij} \sum_k J_{ik}^3 m_k (1-m_k^2) + \frac{2}{3} J_{ij}^3 (1-3m_i^2)(1-3m_j^2) \\ - \delta_{ij} \sum_{k\ell} J_{ik} J_{k\ell} J_{\ell i} (1-m_k^2)(1-m_\ell^2) + 4J_{ij} m_i m_j \sum_k J_{jk} J_{ki} (1-m_k^2) \Bigg]. \tag{4.22}$$

Part II

HUMAN LEARNING AND INFORMATION PROCESSING WITH COMPLEX NETWORKS

While humans connect and interact to form complex social networks, at the individual level humans also communicate and process information using complex networks of interconnected stimuli and concepts. The organization of these systems – from language and music to literature and science – encode the types of information that a person can send and receive. But how do humans uncover the large-scale structures of networks in the world around them? Moreover, how can we quantify the amount of information that a system communicates to a human observer? Here, we answer these questions by combining experimental tools from cognitive science with theoretical concepts from information theory and network science. In Chapter 5, we introduce the emerging field of graph learning, reviewing what is known about how humans infer and internally represent networks. In Chapter 6, we develop a model for how humans uncover the structure of connections between items in a sequence – such as words in a sentence or concepts in a classroom lecture – and we test our model in a series of behavioral experiments. Using the free entropy principle, we demonstrate that mental errors play a crucial role in forming human representations of networks. In Chapter 7, we present a framework for quantifying the amount of information that a network communicates to a human observer. Applying our method to an array of real-world communication networks, we find that they are organized to support the rapid and efficient transmission of information. In combination, these results suggest that mental errors impact how humans learn and perceive networks, and that real-world communication networks are shaped by the pressures of information transmission.

5

HOW HUMANS LEARN AND REPRESENT NETWORKS

This chapter contains work from Lynn, Christopher W., and Danielle S. Bassett. "How humans learn and represent networks." Proceedings of the National Academy of Sciences, in press.

Abstract

Humans receive information from the world around them in sequences of discrete items – from words in language or notes in music to abstract concepts in books and websites on the Internet. In order to model their environment, from a young age people are tasked with learning the network structures formed by these items (nodes) and the connections between them (edges). But how do humans uncover the large-scale structures of networks when they only experience sequences of individual items? Moreover, what do people's internal maps and models of these networks look like? Here, we introduce *graph learning*, a growing and interdisciplinary field studying how humans learn and represent networks in the world around them. Specifically, we review progress toward understanding how people uncover the complex webs of relationships underlying sequences of items. We begin by describing established results showing that humans can detect fine-scale network structure, such as variations in the probabilities of transitions between items. We next present recent experiments that directly control for differences in transition probabilities, demonstrating that human behavior depends critically on the mesoscale and macroscale properties of networks. Finally, we introduce computational models of human graph learning that make testable predictions about the impact of network structure on people's behavior and cognition. Throughout, we highlight open questions in the study of graph learning that will require creative insights from cognitive scientists and network scientists alike.

5.1 INTRODUCTION

Our experience of the world is punctuated by discrete items and events, all connected by a hidden network of forces, causes, and associations. Just as navigation requires a mental map of one's physical surroundings (*262*, *666*), anticipation, planning, perception, and communication all depend on a person's ability to learn the network structure connecting items and events in their environment (*55*, *381*, *537*). For example, in order to identify the boundaries between words, children as young as eight months old identify subtle variations in the network of transitions between syllables in spoken language (*576*). Within their first 30 months, toddlers already learn enough words to

form complex language networks that exhibit robust structural features (*202*, *313*, *617*). By the time we reach adulthood, graph learning enables us to understand and produce language (*227*, *576*), flexibly and adaptively learn words (*349*, *350*), parse continuous streams of stimuli (*576*), build social intuitions (*667*), perform abstract reasoning (*99*), and categorize visual patterns (*217*). In this way, our ability to learn the structures of networks supports a wide range of cognitive functions.

Our capacity to infer and represent complex relationships has also enabled humans to construct an impressive array of networked systems, from language (*121*, *192*, *636*) and music (*407*) to social networks (*51*, *250*), the Internet (*12*, *207*), and the web of concepts that constitute the arts and sciences (*429*, *477*). Moreover, individual differences in cognition, such as those driven by learning disabilities and age, give rise to variations in the types of network structures that people are able to construct (*71*, *195*). Therefore, studying how humans learn and represent networks will not only inform our understanding of how we perform many of our basic cognitive functions, but will also shed light on the structure and function of networks in the world around us.

Here, we provide a brief introduction to the field of graph learning, spanning the experimental techniques and network-based models, theories, and intuitions recently developed to study the effects of network structure on human cognition and behavior. Given the highly interdisciplinary nature of the field – which draws upon experimental methods from cognitive science and linguistics and builds upon computational techniques from network science, information theory, and statistical learning – we aim to present an accessible overview with simple motivating examples.

We focus particular attention on understanding how people uncover the structure of connections between items in a sequence, such as syllables and words in spoken and written language, concepts in books and classroom lectures, or notes in musical progressions. We begin by discussing experimental results demonstrating that humans are adept at detecting differences in the probabilities of transitions between items, and how such transitions connect and combine to form networks that encode the large-scale structure of entire sequences. We then present recent experiments that measure the effects of network structure on human behavior by directly controlling for differences in transition probabilities, followed by a description of the computational models that have been proposed to account for these network effects. We conclude by highlighting some of the open research directions stemming from recent advances in graph learning, including important generalizations of existing graph learning paradigms and direct implications for understanding the structure and function of real-world networks.

5.2 LEARNING TRANSITION PROBABILITIES

As humans navigate their environment and accumulate experience, one of the brain's primary functions is to infer the statistical relationships governing causes and effects (*379*, *709*). Given a sequence of items, perhaps the simplest statistics available to a learner are the frequencies of transitions from one item to another. Naturally, the

field of statistical learning, which is devoted to understanding how humans extract statistical regularities from their environment, has predominantly focused on these simple statistics. For example, consider spoken language, wherein distinct syllables transition from one to another in a continuous stream without pauses or demarcations between words (*107*). How do people segment such continuous streams of data, identifying where one word starts and another begins? The answer, as research has robustly established (*29*, *30*, *564*, *583*), lies in the statistical properties of the transitions between syllables.

The ability to detect words within continuous speech was initially demonstrated by Saffran et al. (*576*), who exposed infants to sequences of four pseudowords, each consisting of three syllables (Fig. 5.1a). The order of syllables within each word remained consistent, yielding a within-word transition probability of 1. However, the order of the words was random, yielding a between-word transition probability of 1/3. Infants were able to reliably detect this difference in syllable transition probabilities, thereby providing a compelling mechanism for word identification during language acquisition. This experimental paradigm has since been generalized to study statistical learning in other domains, with stimuli ranging from colors (*673*) and shapes (*217*) to visual scenes (*101*) and physical actions (*44*). Indeed, the capacity to uncover variations in transition probabilities is now recognized as a central and general feature of human learning (*29*, *30*, *564*, *583*).

5.3 LEARNING NETWORK STRUCTURE

Although individual connections between items provide important information about the structure of a system, they do not tell the whole story. Connections also combine and overlap to form complex webs that characterize the higher-order structure of our environment. To study these structures, scientists have increasingly turned to the language of network science (*478*), conceptualizing items as nodes in a network with edges defining possible connections between them (see Fig. 5.6 for a primer on networks). One can then represent a sequence of items, such as the stream of syllables in spoken language, as a walk through this underlying network (*242*, *351*, *407*, *419*, *584*). This perspective has been particularly useful in the study of artificial grammar learning (*147*, *267*, *550*), wherein human subjects are tasked with inferring the grammar rules (i.e., the network of transitions between letters and words) underlying a fabricated language.

By translating items and connections into the language of network science, one inherits a powerful set of descriptive tools and visualization techniques for characterizing different types of structures. For example, consider once again the statistical learning experiment of Saffran et al. ((*576*); Fig. 5.1a). Simply by visualizing the transition structure as a network (Fig. 5.1b), it becomes clear that the syllables split naturally into four distinct clusters corresponding to the four different words in the artificial language. This observation raises an important question: When parsing words (or performing any other learning task), are people only sensitive to differences in individual connections,

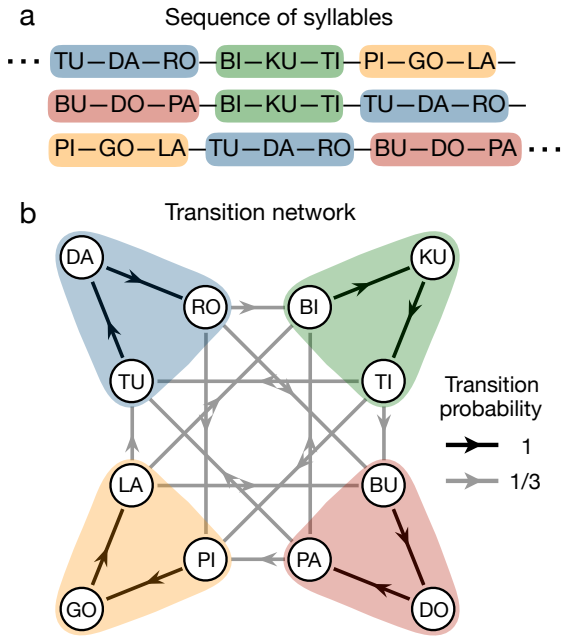


Figure 5.1: Transitions between syllables in the fabricated language of Saffran et al. (*576*). (*a*) A sequence containing four different pseudowords: *tudaro* (blue), *bikuti* (green), *budopa* (red), and *pigola* (yellow). When spoken, the sequence forms a continuous stream of syllables, without clear boundaries between words. The transition probability from one syllable to another is 1 if the transition occurs within a word and 1/3 if the transition occurs between words. This difference in transition probabilities allows infants to segment spoken language into distinct words (*360*, *564*, *576*). (*b*) Transitions between syllables form a network, with edge weights representing syllable transition probabilities. A random walk in the transition network defines a sequence of syllables in the pseudolanguage. The four pseudowords form distinct communities (highlighted) that are easily identifiable by eye. Reprinted from (*360*) with permission from Elsevier.

or do they also uncover large-scale features of the underlying network? In what follows, we describe recent advances in graph learning that shed light on precisely this question.

5.3.1 *Learning local structure*

The simplest properties of a network are those corresponding to individual nodes and edges, such as the weight of an edge, which determines the strength of the connection between two nodes, and the degree of a node, or its number of connections. For example, edge weights can represent transition probabilities between syllables or words (*29*, *30*, *564*, *583*), similarities between different semantic concepts (*55*, *636*), or strengths of social interactions (*51*, *250*). Meanwhile, significant effort has focused on understanding how humans learn the network structure surrounding individual nodes (*8*, *45*, *123*, *133*, *202*, *260*, *716*). For example, the degree defines the connectedness of a node, such as the number of links pointing to a website (*12*, *207*, *432*), the number of friends that a person has (*51*), or the number of citations accumulated by a scientific

paper (*429*). Notably, many of the networks that people encounter on a daily basis – including language, social, and hyperlink networks – exhibit heavy-tailed degree distributions, with many nodes of low degree and a select number of high-degree hubs (*50*, *55*, *96*, *121*, *192*, *207*, *432*, *477*, *636*).

Significant research has now demonstrated that people are able to learn the local network properties of individual nodes and edges, such as the transition probabilities between syllables in the previous section (*29*, *30*, *564*, *583*). To illustrate the impact of network structure on human behavior, we consider a recently-developed experimental paradigm (*351*, *419*), while noting that similar results have also been achieved using variations on this approach (*242*, *360–362*, *584*, *667*). Specifically, each subject is shown a sequence of stimuli, with the order of stimuli defined by a random walk on an underlying transition network (Fig. 5.2a). Subjects are asked to respond to each stimulus by performing an action (and to avoid confounds the assignment of stimuli to nodes in the network is randomized across subjects). By measuring the speed with which subjects respond to stimuli, one can infer their expectations about the network structure: A fast reaction reflects a strongly-anticipated transition, while a slow reaction reflects a weakly-anticipated (or surprising) transition (*330*, *351*, *419*, *434*).

Intuitively, one should expect a subject's anticipation to increase (and thus their reaction time to decrease) for edges representing more probable transitions. In order to test this prediction, we note that for a random walk in an unweighted and undirected network, the transition probability from one node i to a neighboring node j is given by $P_{ij} = 1/k_i$, where k_i is the degree of node i. Aligning with intuition, researchers have shown that people's reaction times are positively correlated with the degree of the previous stimulus (Fig. 5.2b), and therefore, people are better able to anticipate more probable transitions (*351*, *419*). Interestingly, significant research has also established similar results in language networks, with people reading words more quickly if they occur more frequently or appear in more contexts (*8*, *45*, *221*). Conversely, humans tend to slow down and produce more errors when attempting to recall words with a large number of semantic associations, a phenomenon known as the fan effect (*21*, *22*). Together, these results demonstrate that humans are sensitive to variations in the local properties of individual nodes and edges, but what about the mesoscale and macroscale properties of a network?

5.3.2 *Learning mesoscale structure*

The mesoscale structure of a network reflects the organizational properties of groups of nodes and edges. One such property is clustering, or the tendency for a pair of nodes with a common neighbor to form a connection themselves. This tendency is clearly observed in social networks, where people with a common friend are themselves more likely to become friends. Similar principles govern the mesoscale structure of many other real-world networks, with items such as words, scientific papers, and webpages all exhibiting high clustering (*429*, *461*, *613*, *699*). As nodes cluster together, they often give rise to a second mesoscale property – modular structure – which is characterized

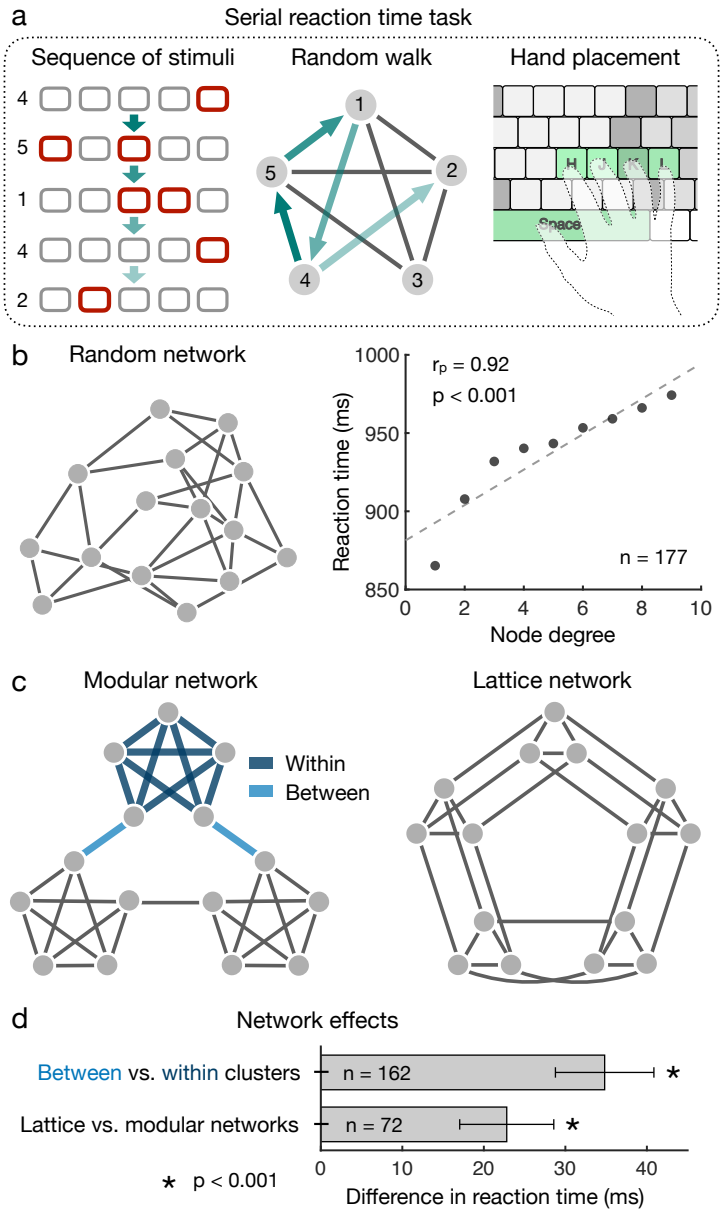


Figure 5.2: Human behavior depends on network topology. (*a*) We consider a serial reaction time experiment in which subjects are shown sequences of stimuli and are asked to respond by performing an action. Here, each stimulus consists of five squares, one or two of which are highlighted in red (left); the order of stimuli is determined by a random walk on an underlying network (center); and for each stimulus, the subject presses the keys on the keyboard corresponding to the highlighted squares (right). (*b*) Considering Erdös-Rényi random transition networks with 15 nodes and 30 edges (left), subjects' average reaction times to a transition $i \to j$ increase as the degree k_i of the preceding node increases (right). Equivalently, subjects' reaction times increase as the transition probability $P_{ij} = 1/k_i$ decreases (*419*). (*c*) To control for variations in transition probabilities, we consider two networks with constant degree $k = 4$: a *modular network* consisting of three communities of five nodes each (left) and a *lattice network* representing a 3×5 grid with periodic boundary conditions (right). (*d*) Experiments indicate two consistent effects of network structure. First, in the modular network, reaction times for between-cluster transitions are longer than for within-cluster transitions (*351*, *361*, *362*, *419*). Second, reaction times are longer on average for the lattice network than for the modular network (*351*, *419*).

by tightly-connected modules or communities of nodes. Such modular structure is now recognized as a ubiquitous feature of networks in our environment (*479*), with language splitting into groups of semantically or phonetically similar words (*99*, *121*), people forming social cliques (*51*, *250*, *455*), and websites clustering into online communities (*207*).

Over the past ten years, researchers have made signifiant strides toward understanding how the mesoscale properties of a network impact human learning and behavior. Words with higher clustering are more likely to be acquired during language learning (*260*), while words with lower clustering are easier to recognize in long-term memory (*688*) and convey processing (*133*, *716*) and production (*134*) benefits. Additionally, in a series of cognitive and neuroimaging experiments, researchers have found that a network's modular structure has a significant impact on human behavior and neural activity. For example, people are able to detect the boundaries between communities in a network just by observing sequences of nodes (*351*, *361*, *362*, *419*, *584*). Moreover, strong modular structure helps people build more accurate mental representations of a network, thereby allowing humans to better anticipate future items and events (*351*, *361*, *362*, *419*, *584*).

5.3.3 *Learning global structure*

In addition to their local and mesoscale features, networks also have global properties that depend on the entire architecture of nodes and edges. Perhaps the most well-studied global property is small-world structure, wherein each node connects to every other node in only a small number of steps (*699*). Small-world topology has been observed in an array of networks that humans are tasked with learning, including social relationships (*375*), web hyperlinks (*12*), scientific citations (*429*), and semantic associations in language (*96*, *121*). Moreover, in a particularly compelling example of the relationship between global network structure and human cognition, the small-world structure of people's learned language networks has been shown to vary from person to person, decreasing with age (*195*) and in people with learning disabilities (*71*).

While small-worldness describes the structure of an entire network, there are also measures that relate individual nodes to a network's global topology, including centrality (a measure of a node's role in mediating long-distance connections), communicability (a measure of the number of paths connecting a pair of nodes), and coreness (a measure of how deeply embedded a node is in a network). Global measures such as these have recently been shown to impact human learning and cognition, indicating that humans are sensitive to the global structure of networks in their environment. For example, in the reaction time experiments described above (Fig. 5.2a), people responded more quickly, and therefore were better able to anticipate, nodes with low centrality (*351*). In a related experiment, neural activity was shown to reflect the communicability between pairs of stimuli in an underlying transition network (*242*). Finally, as children learn language, they more readily acquire and produce words with low coreness (*123*). Together, these results point to a robust and general relationship between large-scale

network structure and human cognition. However, might these large-scale network effects simply be driven by confounding variations in the local network structure?

5.3.4 *Controlling for differences in local structure*

To disentangle the effects of large-scale network structure from those of local structure, recent research has directly controlled for differences in transition probabilities by focusing on specific families of networks (*351*, *361*, *419*, *584*). Recall that for random walks on unweighted, undirected networks, the transition probabilities are determined by node degrees. Therefore, to ensure that all transitions have equal probability, one can simply focus on graphs with constant degree but varying topology. For example, consider the modular and lattice graphs shown in Fig. 5.2c. Since both networks have constant degree 4 (and therefore constant transition probability $1/4$ across all edges), any variation in behavior or cognition between different parts of a network, or between the two networks themselves, must stem from the networks' global topologies.

This approach was first developed by Schapiro et al. (*584*), who demonstrated that people are able to detect the transitions between clusters in the modular graph (Fig. 5.2c), and that these between-cluster transitions yield distinct patterns of neural activity relative to within-cluster transitions. Returning to the reaction time experiment (Fig. 5.2a), it was shown that subjects react more quickly to (and therefore are able to better anticipate) within-cluster transitions than between-cluster transitions ((*351*, *419*); Fig 5.2d). Moreover, people exhibit an overall decrease in reaction times for the modular graph relative to the lattice graph ((*351*, *419*); Fig. 5.2d).

These results, combined with findings in similar experiments (*361*, *362*), demonstrate that humans are sensitive to features of mesoscale and global network topology, even after controlling for differences in local structure. Thus, not only are humans able to learn individual transition probabilities, as originally demonstrated in seminal statistical learning experiments (Fig. 5.1), they are also capable of uncovering some of the complex structures found in our environment. But how do people learn the large-scale features of networks from past observations?

5.4 MODELING HUMAN GRAPH LEARNING

Experiments spanning cognitive science, neuroscience, linguistics, and statistical learning have established that human behavior and cognition depend on the mesoscale and global topologies of networks in their environment. To understand how people detect these global features, and to make quantitative predictions about human behavior, one requires computational models of how humans construct internal representations of networks from past experiences. Here, we again focus on understanding how people learn the networks of transitions underlying observed sequences of items, such as words in a sentence, concepts in a book or classroom lecture, or notes in a musical

progression. Interestingly, humans systematically deviate from the most accurate, and perhaps the simplest, learning rule.

To make these ideas concrete, consider a sequence of items described by the transition probability matrix P, where P_{ij} represents the conditional probability of one item i transitioning to another item j. Given an observed sequence of items, one can imagine estimating P_{ij} by simply dividing the number of times i has transitioned to j (denoted by n_{ij}) by the number of times i has appeared (which equals $\sum_k n_{ik}$):

$$\hat{P}_{ij} = \frac{n_{ij}}{\sum_k n_{ik}}. \tag{5.1}$$

In fact, not only is this perhaps the simplest estimate one could perform, it is also the most accurate (or maximum likelihood) estimate of the transition probabilities from past observations (*93*). An important feature of maximum likelihood estimation is that it gives an unbiased approximation of the true transition probabilities; that is, the estimated transition probabilities $\hat{P}_{ij}$ are evenly distributed about their true values P_{ij}, independent of the large-scale structure of the network (*93*). However, we have seen that people's behavior and cognition depend systematically on mesoscale and global network properties, even when transition probabilities are held constant (*242*, *351*, *361*, *419*, *584*). Thus, when constructing internal representations, humans allow higher-order network structure to influence their estimates of individual transition probabilities, thereby deviating from maximum likelihood estimation (*419*).

To understand the impact of network topology on human cognition, researchers have recently proposed a number of models describing how humans learn and represent transition networks (*23*, *180*, *242*, *324*, *419*, *444*, *445*, *451*). Notably, many of these models share a common underlying mechanism: that instead of just counting transitions of length one (as in maximum likelihood estimation), humans also include transitions of lengths two, three, or more in their representations (*23*, *242*, *419*, *445*, *451*, *480*). Mathematically, by combining transitions of different distances, the estimated transition probabilities take the form

$$\hat{P}_{ij} = C \sum_{t \geqslant 1} f(t) n_{ij}^{(t)}, \tag{5.2}$$

where $n_{ij}^{(t)}$ represents the number of times that i has transitioned to j in t steps, $f(t)$ defines the weight placed on transitions of a given distance, and C is a normalization constant. Interestingly, this simple prediction can be derived from a number of different cognitive theories – including the temporal context model of episodic memory (*324*), temporal difference learning and the successor representation in reinforcement learning (*170*, *247*, *646*), and the free energy principle from information theory (*419*). But how does combining transitions over different distances allow people to learn the structure of a network?

To answer this question, it helps to consider different choices for the function $f(t)$. Typically, $f(t)$ is assumed to be decreasing such that longer-distance associations contribute more weakly to a person's network representation (*247*, *419*, *646*). If $f(t)$ is

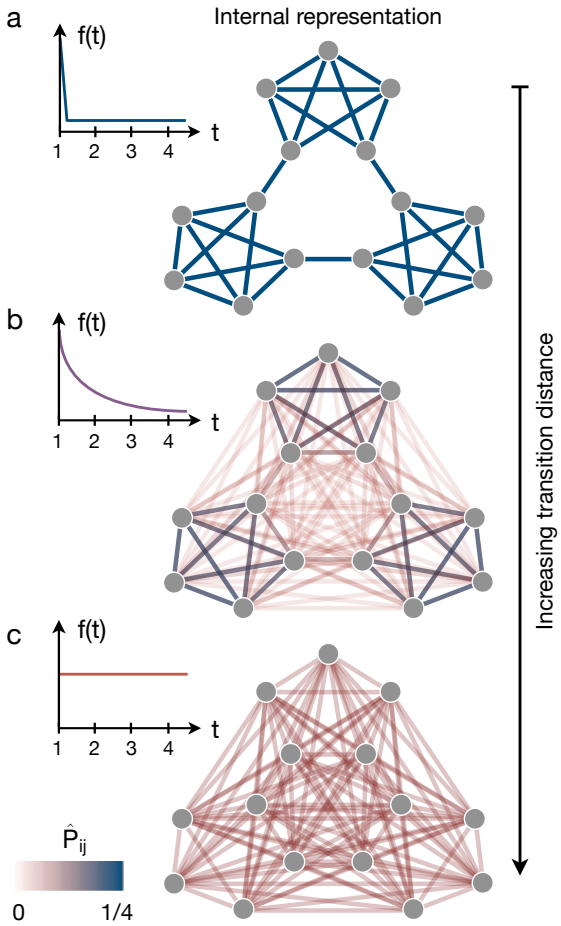


Figure 5.3: Mesoscale and global network features emerge from long-distance associations. (*a*) Illustration of the weight function $f(t)$ (left) and the learned network representation $\hat{P}$ for learners that only consider transitions of length one. The estimated structure resembles the true modular network. (*b*) For learners that down-weight transitions of longer distances, higher-order features of the transition network, such as community structure, organically come into focus, yielding higher expected probabilities for within-cluster transitions than for between-cluster transitions. (*c*) For learners that equally weigh transitions of all distances, the internal representation becomes all-to-all, losing any resemblance to the true transition network. Panels a-c correspond to learners that include progressively longer transitions in their network estimates. Adapted from (*419*).

a delta function centered at $t = 1$ (Fig. 5.3a), then the learner focuses on transitions of length one. In this case, people simply perform maximum likelihood estimation, resulting in an unbiased estimate of the true transition structure P. Conversely, if $f(t)$ is uniform over all time scales $t \geqslant 1$, then the learner equally weighs transitions of all distances (Fig. 5.3c), and the estimate $\hat{P}$ loses any resemblance to the true transition structure P. Importantly, however, for learners who combine transitions over intermediate distances (Fig. 5.3b), we find that large-scale features of the network organically come into focus. Consider, for example, the modular network from Fig. 5.2c.

By combining transitions of lengths two, three, or more, humans tend to over-weigh the associations within communities and under-weigh the transitions between communities (Fig. 5.3b). This simple observation explains why people are surprised by cross-cluster transitions ((*351*, *419*); Fig. 5.2d), why sequences in lattice and random networks are more difficult to anticipate ((*351*, *419*); Fig. 5.2d), and how people detect the boundaries between clusters (*361*, *362*, *584*).

More generally, the capacity to learn the large-scale structure of a network enables people to perform many basic cognitive functions, from anticipating non-adjacent dependencies between syllables and words (*15*, *480*) to planning for future events (*7*, *31*) and estimating future rewards (*247*, *646*). Using models similar to that above, researchers have been able to predict the impacts of network structure on human behavior in reinforcement learning tasks (*451*), pattern detection in random sequences (*23*, *445*), and variations in neural activity (*242*, *445*, *584*). Notably, the explained effects span various types of behavioral and neural observations, including reaction times (*329*, *351*, *419*), data segmentation (*361*, *362*, *584*), task errors (*351*, *419*), randomness detection (*213*), EEG signals (*628*), and fMRI recordings (*329*, *584*). Together, these results indicate that people's ability to detect the mesoscale and global structure of a network emerges not just from their capacity to learn individual edges, but also from their capacity to associate items across spatial, temporal, and topological scales.

5.5 THE FUTURE OF GRAPH LEARNING

Past and current advances in graph learning inspire new research questions at the intersection of cognitive science, neuroscience, and network science. Here, we highlight a number of important directions, beginning with possible generalizations of the existing graph learning paradigm before discussing the implications of graph learning for our understanding of the structures and functions of real-world transition networks.

5.5.1 *Extending the graph learning paradigm*

Most graph learning experiments, including those discussed in Figs. 5.1 and 5.2, present each subject with a sequence of stimuli defined by a random walk on a (possibly weighted and directed) transition network (*147*, *242*, *267*, *351*, *360–362*, *419*, *480*, *550*, *576*, *584*, *667*). Equivalently, in the language of stochastic processes, each sequence represents a stationary Markov process (*567*). Although random walks offer a natural starting point in the study of graph learning, they are also constrained by three main assumptions: (i) that the underlying transition structure remains static over time (stationarity), (ii) that future stimuli only depend on the current stimulus (the Markov property), and (iii) that the sequence is predetermined without input from the observer. Future graph learning experiments can test the boundaries of these constraints by systematically generalizing the existing graph learning paradigm.

5.5.1.1 *Stationarity*

While most graph learning experiments focus on static transition networks, many of the networks that humans encounter in the real world either evolve in time or overlap with other networks in the environment (*51*, *71*, *192*, *617*, *636*). Therefore, rather than simply investigating people's ability to learn a single network, future experiments should study the capacity for humans to detect the dynamical features of an evolving network (Fig. 5.4a) or differentiate the distinct features of multiple networks. Early results indicate that, when observing a sequence of stimuli that shifts from one transition structure to another, people's learned representation of the first network influences their behavior in response to the second network, but that these effects diminish with time (*351*). This gradual "unlearning" of network structure raises an important question for future research: Rather than investigating how network properties facilitate learning – as has been the focus of most graph learning studies – can we determine which properties make a network difficult to forget?

5.5.1.2 *The Markov property*

Thus far, in keeping with the majority of existing graph learning research, we have focused exclusively on sequences in which the next stimulus depends only on the current stimulus; that is, we have focused on sequences that obey the Markov property (*567*). However, almost all sequences of stimuli or items in the real world involve long-range correlations and dependencies (Fig. 5.4b). For example, the probability of a word in spoken language depends not just on the previous word, but also the earlier words in the sentence and the broader context in which the sentence exists (*18*). Similarly, musical systems often enforce constraints on the length and structure of sequences, thereby inducing long-range dependencies between notes (*337*). Interestingly, given mounting evidence that people construct long-distance associations (*23*, *242*, *419*, *445*, *451*, *480*), the resulting internal estimates of transition structures resemble non-Markov processes (*419*). Therefore, future research could investigate whether the learning long-distance associations enables people to infer the non-Markov features of sequences in daily life.

5.5.1.3 *Information seeking*

Finally, although many of the sequences that humans observe are prescribed without input from the observer, there are also settings in which people have agency in determining the structure of a sequence. For example, when surfing the Internet (*6*, *189*, *487*, *702*) or following a trail of scientific citations (*429*), people choose their paths through the underlying hyperlink and citation networks. In this way, people are able to seek out information about networks structures rather than simply having the information presented to them (Fig. 5.4c). Such information seeking has been shown to vary by person (*487*) and to depend crucially on the topology of the underlying network (*6*, *189*, *702*). Moreover, when retrieving information from memory, humans search through

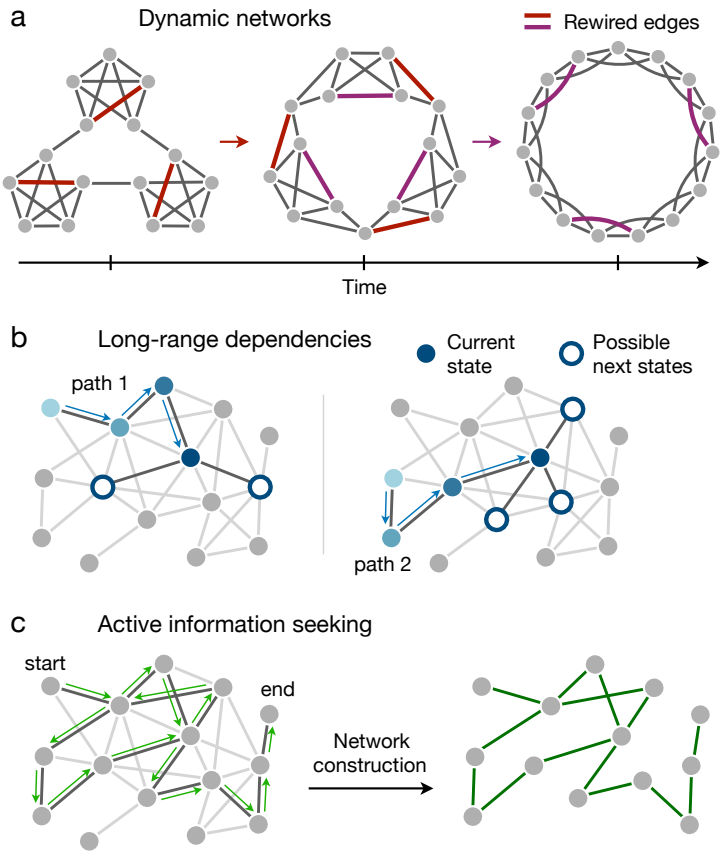


Figure 5.4: Generalizations of the graph learning paradigm. (*a*) Transition networks often shift and change over time. Such non-stationary transition probabilities can be described using dynamical transition networks, which evolve from one network (for example, the modular network on the left) to another (for example, the ring network on the right) by iteratively rewiring edges. (*b*) Many real-world sequences have long-range dependencies, such that the next state depends not just on the current state, but also on a number of previous states (*18*, *337*). For example, path 1 in the displayed network yields two possibilities for the next state (left), while path 2 yields a different set of three possible states (right). (*c*) Humans often actively seek out information by choosing their path through a transition network, rather than simply being presented with a prescribed sequence. Such information seeking yields a subnetwork containing the nodes and edges traversed by the walker.

their stored networks of associations (*542*), often performing search strategies that resemble optimal foraging in physical space (*37*, *314*, *345*). In the context of graph learning, allowing subjects to actively seek information raises a number of compelling questions: Does choosing their path through a transition network enable subjects to more efficiently learn its topology? Or does the ability to seek information lead people to form biased representations of the true transition structure (*344*, *595*)? These questions, combined with the directions described above, highlight some of the exciting extensions of graph learning that will require creative insights and collaborative contributions from cognitive scientists and network scientists alike.

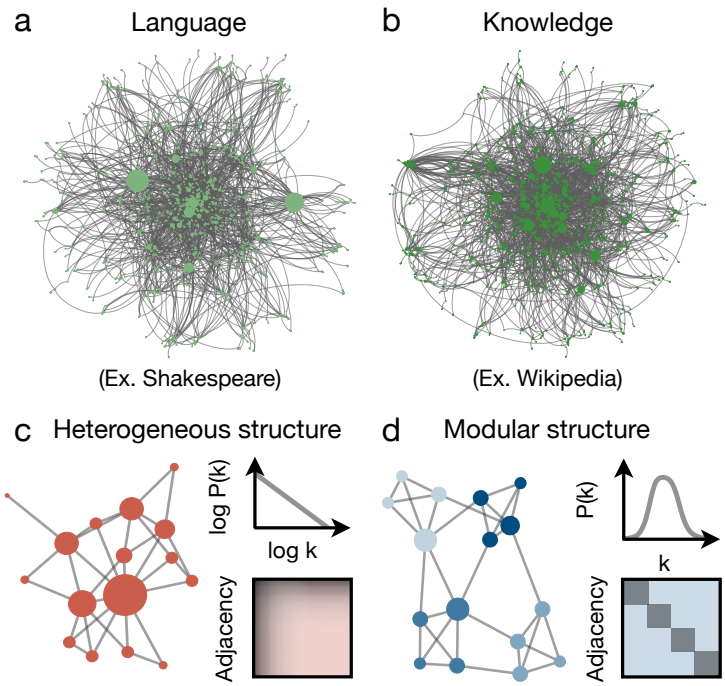


Figure 5.5: Real transition networks exhibit hierarchical structure. (*a*) A language network constructed from the words (nodes) and transitions between them (edges) in the complete works of Shakespeare. (*b*) A knowledge network of hyperlinks between pages on Wikipedia. (*c*, *d*) Many real-world transition networks exhibit hierarchical organization (*547*), which is characterized by two topological features: (*c*) Heterogeneous structure, which is often associated with scale-free networks, is typically characterized by a power-law degree distribution and the presence of high-degree hub nodes (*50*). (*d*) Modular structure is defined by the presence of clusters of nodes with dense within-cluster connectivity and sparse between-cluster connectivity (*250*).

5.5.2 *Studying the structure of real-world networks*

In addition to shedding light on human behavior and cognition, the study of graph learning also has the promise to offer insights into the structure and function of real-world networks. Indeed, there exists an intimate connection between human cognition and networks: While people rely on networked systems to perform a wide range of tasks, from communicating using language (Fig. 5.5a) and music to storing and retrieving information through science and the Internet (Fig. 5.5b), many of these networks have evolved with or were explicitly designed by humans. Therefore, just as humans are adept at learning the structure of networks, one might suspect that some networks are structured to support human learning and cognition.

The perspective that cognition may constrain network structure has recently shed light on the organizational properties of some real-world networks (*55*, *96*), including the small-world structure and power-law degree distributions exhibited by semantic and word co-occurrence networks (*121*, *192*, *636*), and the scale-free structure of the

connections between concepts on Wikipedia (*432*). Interestingly, many of the networks with which humans interact share two distinct structural features: (i) They are heterogeneous (Fig. 5.5c), characterized by the presence of hub nodes with unusually high degree (*50*, *96*, *121*, *477*, *636*), and (ii) they are modular (Fig. 5.5d), characterized by the existence of tightly-connected clusters (*96*, *207*, *250*, *461*, *636*). Together, heterogeneity and modularity represent the two defining features of hierarchical organization, which has now been observed in a wide array of man-made networks (*26*, *547*). Could it be that the shared structural properties of these networks arise from their common functional purpose: to facilitate human learning and communication?

Graph learning provides quantitative models and experimental tools to begin answering questions such as these (*420*). For example, experimental results, such as those discussed in Fig. 5.2, indicate that modular structure improves people's ability to anticipate transitions (*351*, *419*), and this result has been confirmed numerically using models of the form in Fig. 5.3 (*419*). Moreover, the high-degree hubs found in heterogeneous networks have been shown to help people search for information (*6*, *702*). Together, these results demonstrate that graph learning offers a unique and constructive lens through which to study networks in the world around us.

5.6 CONCLUSIONS AND OUTLOOK

Understanding how people learn and represent the complex relationships governing their environment remains one of the greatest open problems in the study of human cognition. On the heels of decades of research in cognitive science and statistical learning investigating how humans detect the local properties of individual items and the connections between them (*29*, *30*, *44*, *101*, *217*, *564*, *576*, *583*, *673*), conclusive evidence now demonstrates that human behavior, cognition, and neural activity depend critically on the large-scale structure of items and connections (*242*, *351*, *360–362*, *419*, *584*, *667*). By casting the items and connections in our environment as nodes and edges in a network, scientists can now explore the impact of network structure on human cognition in a unified and principled framework.

Although the experimental and numerical foundation of the field has been laid, graph learning remains a budding area of research offering a wealth of interdisciplinary opportunities. From new cognitive modeling techniques (Fig. 5.3) and extensions of existing experimental paradigms (Fig. 5.4) to novel applications in the study of real-world networks (Fig. 5.5), graph learning is primed to alter the way we think about human cognition, complex networks, and the myriad ways in which they intersect.

5.7 SUPPLEMENTARY MATERIAL

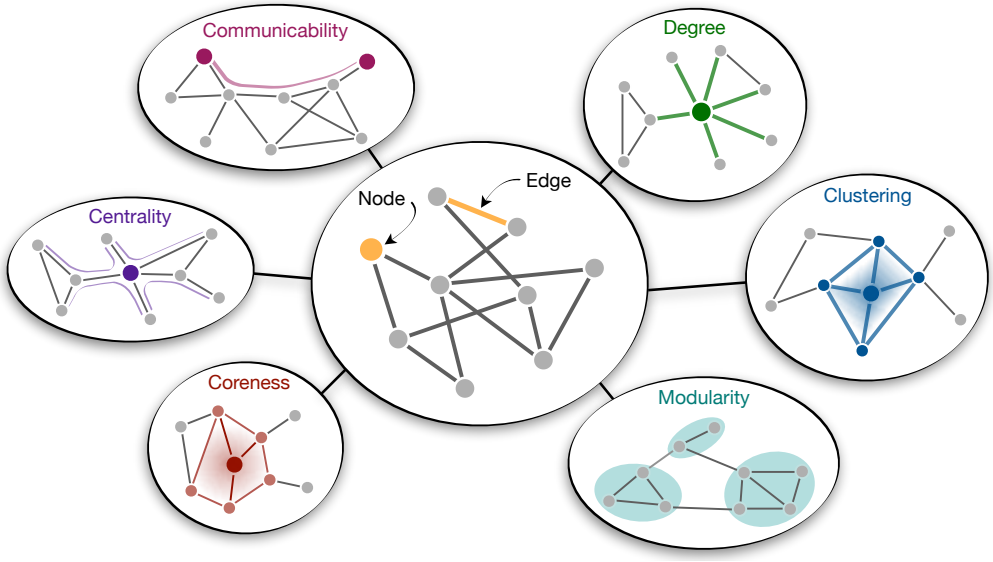


Figure 5.6: A primer on network properties. (*Center*) Nodes, illustrated by circles, represent stimuli, items, or states in a sequence. Edges, illustrated by lines, connect pairs of nodes if it is possible to transition from one node to the other. The organization of edges among nodes is referred to as the network's *topology* or *structure*. (*Circumjacent*) A network's topology can be described using properties that characterize its local, mesoscale, or global organization. For example, the simplest local property is the degree of a node (green), or the number of edges emanating from a node. Two notions of mesoscale structure include (i) the clustering coefficient (blue), or the ratio of connected triangles to connected triples of nodes, and (ii) modularity (turquoise), where there exist communities of nodes with internally dense and externally sparse connections. Finally, global measures include (i) coreness (red), or the ability of a node to withstand the removal of nodes with low degree, (ii) notions of centrality (purple) such as betweenness centrality, which quantifies the importance of a node for facilitating long-distance connections, and (iii) communicability (magenta), which captures the number of paths of various lengths connecting two nodes. Collectively, the network representation and associated properties can provide critical insights into the structure of the system under study.

6

ABSTRACT REPRESENTATIONS OF EVENTS ARISE FROM MENTAL ERRORS IN LEARNING AND MEMORY

This chapter contains work from Lynn, Christopher W., Ari E. Kahn, Nathaniel Nyema, and Danielle S. Bassett. "Abstract representations of events arise from mental errors in learning and memory." Nature Communications, in press.

Abstract

Humans are adept at uncovering abstract associations in the world around them, yet the underlying mechanisms remain poorly understood. Intuitively, learning the higher-order structure of statistical relationships should involve complex mental processes. Here we propose an alternative perspective: that higher-order associations instead arise from natural errors in learning and memory. Using the free energy principle, which bridges information theory and Bayesian inference, we derive a maximum entropy model of people's internal representations of the transitions between stimuli. Importantly, our model (i) affords a concise analytic form, (ii) qualitatively explains the effects of transition network structure on human expectations, and (iii) quantitatively predicts human reaction times in probabilistic sequential motor tasks. Together, these results suggest that mental errors influence our abstract representations of the world in significant and predictable ways, with direct implications for the study and design of optimally learnable information sources.

6.1 INTRODUCTION

Our experience of the world is punctuated in time by discrete events, all connected by an architecture of hidden forces and causes. In order to form expectations about the future, one of the brain's primary functions is to infer the statistical structure underlying past experiences (*330*, *343*, *635*). In fact, even within the first year of life, infants reliably detect the frequency with which one phoneme follows another in spoken language (*576*). By the time we reach adulthood, uncovering statistical relationships between items and events enables us to perform abstract reasoning (*99*), identify visual patterns (*217*), produce language (*227*), develop social intuition (*667*), and segment continuous streams of data into self-similar parcels (*554*). Notably, each of these functions requires the brain to identify statistical regularities across a range of scales. It has long been known that people are sensitive to differences in individual transition probabilities such as those between words or concepts (*217*, *576*). Additionally, mounting evidence suggests

that humans can also infer abstract (or higher-order) statistical structures, including hierarchical patterns within sequences of stimuli (*444*), temporal regularities on both global and local scales (*180*), abstract concepts within webs of semantic relationships (*524*), and general features of sparse data (*658*).

To study this wide range of statistical structures in a unified framework, scientists have increasingly employed the language of network science (*478*), wherein stimuli or states are conceptualized as nodes in a graph with edges or connections representing possible transitions between them. In this way, a sequence of stimuli often reflects a random walk along an underlying transition network (*242*, *266*, *480*), and we can begin to ask which network features give rise to variations in human learning and behavior. This perspective has been particularly useful, for example, in the study of artificial grammars (*147*), wherein human subjects are tasked with inferring the grammar rules (i.e., the network of transitions between letters and words) underlying a fabricated language (*267*). Complementary research in statistical learning has demonstrated that modules (i.e., communities of densely-connected nodes) within transition networks are reflected in brain imaging data (*584*) and give rise to stark shifts in human reaction times (*360*). Together, these efforts have culminated in a general realization that people's internal representations of a transition structure are strongly influenced by its higher-order organization (*351*, *362*). But how does the brain learn these abstract network features? Does the inference of higher-order relationships require sophisticated hierarchical learning algorithms? Or instead, do natural errors in cognition yield a "blurry" representation, making the coarse-grained architecture readily apparent?

To answer these questions, here we propose a single driving hypothesis: that when building models of the world, the brain is finely-tuned to maximize accuracy while simultaneously minimizing computational complexity. Generally, this assumption stems from a rich history exploring the trade-off between brain function and computational cost (*174*, *674*), from sparse coding principles at the neuronal level (*686*) to the competition between information integration and segregation at the whole-brain level (*669*) to the notion of exploration versus exploitation (*150*) and the speed-accuracy trade-off (*708*) at the behavioral level. To formalize our hypothesis, we employ the free energy principle (*339*), which has become increasingly utilized to investigate constraints on cognitive functioning (*493*) and explain how biological systems maintain efficient representations of the world around them (*232*). Despite this thorough treatment of the accuracy-complexity trade-off in neuroscience and psychology, the prevailing intuition in statistical learning maintains that the brain is either optimized to perform Bayesian inference (*524*, *658*), which is inherently error free, or hierarchical learning (*147*, *180*, *444*, *480*), which typically entails increased rather than decreased computational complexity.

Here, we show that the competition between accuracy and computational complexity leads to a maximum entropy (or minimum complexity) model of people's internal representations of events (*339*, *603*). As we decrease the complexity of our model, allowing mental errors to take effect, higher-order features of the transition network organically come into focus while the fine-scale structure fades away, thus providing a concise mechanism explaining how people infer abstract statistical relationships. To a

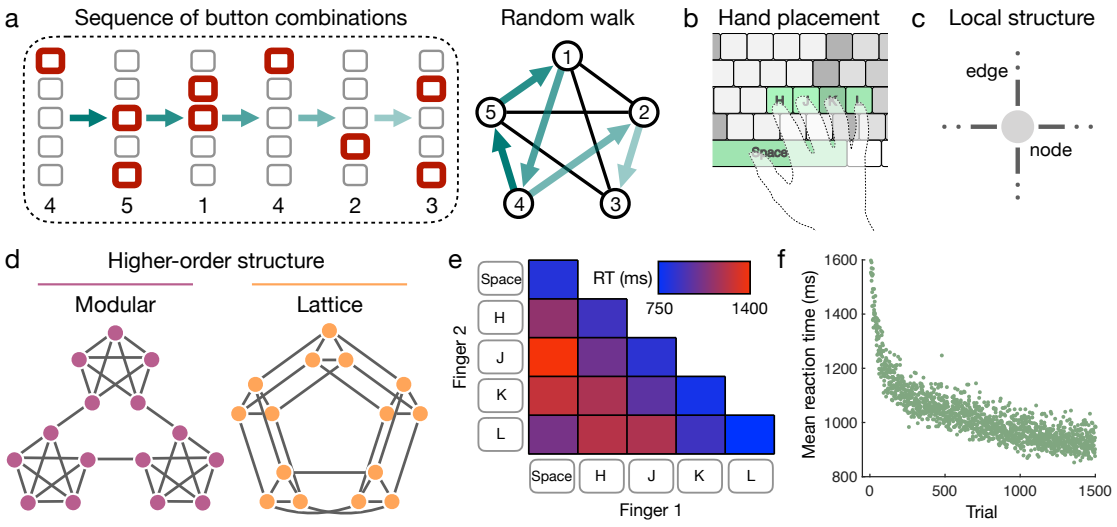


Figure 6.1: Subjects respond to sequences of stimuli drawn as a random walk on an underlying transition graph. (*a*) Example sequence of visual stimuli (left) representing a random walk on an underlying transition network (right). (*b*) For each stimulus, subjects are asked to respond by pressing a combination of one or two buttons on a keyboard. (*c*) Each of the 15 possible button combinations corresponds to a node in the transition network. We only consider networks with nodes of uniform degree $k = 4$ and edges with uniform transition probability 0.25. (*d*) Subjects were asked to respond to sequences of 1500 such nodes drawn from two different transition architectures: a modular graph (left) and a lattice graph (right). (*e*) Average reaction times for the different button combinations, where the diagonal elements represent single-button presses and the off-diagonal elements represent two-button presses. (*f*) Average reaction times as a function of trial number, characterized by a steep drop-off in the first 500 trials followed by a gradual decline in the remaining 1000 trials. In (*e*) and (*f*), averages are taken over responses during random walks on the modular and lattice graphs. Source data are provided as a Source Data file.

broad audience, our model provides an accessible mapping from transition networks to human behaviors, with particular relevance for the study and design of optimally learnable transition structures – either between words in spoken and written language (*147*, *267*, *603*), notes in music (*111*), or even concepts in classroom lectures (*386*).

6.2 RESULTS

6.2.1 *Network effects on human expectations*

In the cognitive sciences, mounting evidence suggests that human expectations depend critically on the higher-order features of transition networks (*266*, *480*). Here, we make this notion concrete with empirical evidence for higher-order network effects in a probabilistic sequential response task (*351*). Specifically, we presented human subjects with sequences of stimuli on a computer screen, each stimulus depicting a row of five grey squares with one or two of the squares highlighted in red (Fig. 7.1a). In response

to each stimulus, subjects were asked to press one or two computer keys mirroring the highlighted squares (Fig. 7.1b). Each of the 15 different stimuli represented a node in an underlying transition network, upon which a random walk stipulated the sequential order of stimuli (Fig. 7.1a). By measuring the speed with which a subject responded to each stimulus, we were able to infer their expectations about the transition structure: a fast reaction reflected a strongly-anticipated transition, while a slow reaction reflected a weakly-anticipated (or surprising) transition (*330*, *351*, *434*, *635*).

While it has long been known that humans can detect differences in transition probabilities – for instance, rare transitions lead to sharp increases in reaction times (*217*, *576*) – more recently it has become clear that people's expectations also reflect the higher-order architecture of transition networks (*351*, *360*, *361*, *584*). To clearly study these higher-order effects without the confounding influence of variations in transition probabilities, here we only consider transition graphs with a uniform transition probability of 0.25 on each edge, thereby requiring nodes to have uniform degree $k = 4$ (Fig. 7.1c). Specifically, we consider two different graph topologies: a *modular* graph with three communities of five densely-connected nodes and a *lattice* graph representing a 3×5 grid with periodic boundary conditions (Fig. 7.1d). Since all transitions across both graphs have uniform probability, any systematic variations in behavior between different parts of a graph, or between the two graphs themselves, must stem from differences in the graphs' higher-order modular or lattice structures.

Regressing out the dependence of reaction times on the different button combinations (Fig. 7.1e), the natural quickening of reactions with time (*39*) (Fig. 7.1f), and the impact of stimulus recency (see Methods), we identify two effects of higher-order network structure on subjects' reactions. First, in the modular graph we find that reactions corresponding to within-cluster transitions are 35 ms faster than reactions to between-cluster transitions ($p < 0.001$, F-test; Tab. 7.1), an effect known as the *cross-cluster surprisal* (*351*, *361*) (Fig. 6.2a). Similarly, we find that people are more likely to respond correctly for within-cluster transitions than between-cluster transitions (Tab. 7.8). Second, across all transitions within each network, we find that reactions in the modular graph are 23 ms faster than those in the lattice graph ($p < 0.001$, F-test; Tab. 7.2), a phenomenon that we coin the *modular-lattice effect* (Fig. 6.2b).

Thus far, we have assumed that variations in human behavior stem from people's internal expectations about the network structure. However, it is important to consider the possible confound of stimulus recency: the tendency for people to respond more quickly to stimuli that have appeared more recently (*41*, *465*). To ensure that the observed network effects are not simply driven by recency, we performed a separate experiment that controlled for recency in the modular graph by presenting subjects with sequences of stimuli drawn according to Hamiltonian walks, which visit each node exactly once (*584*). Within the Hamiltonian walks, we still detect a significant cross-cluster surprisal effect (Fig. 6.2a; Tabs. 7.3, 7.4, and 7.5). Additionally, we controlled for recency in our initial random walk experiments by focusing on stimuli that previously appeared a specific number of trials in the past. Within these recency-controlled data, we find that both the cross-cluster surprisal and modular-lattice effects remain significant

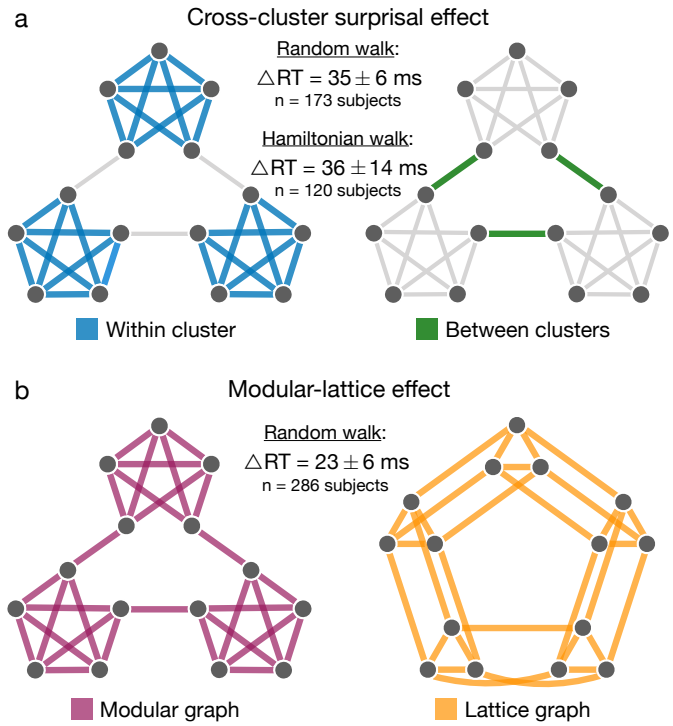


Figure 6.2: The effects of higher-order network structure on human reaction times. (*a*) Cross-cluster surprisal effect in the modular graph, defined by an average increase in reaction times for between-cluster transitions (right) relative to within-cluster transitions (left). We detect significant differences in reaction times for random walks ($p < 0.001$, $t = 5.77$, $df = 1.61 \times 10^5$) and Hamiltonian walks ($p = 0.010$, $t = 2.59$, $df = 1.31 \times 10^4$). For the mixed effects models used to estimate these effects, see Tabs. 7.1 and 7.3. (*b*) Modular-lattice effect, characterized by an overall increase in reaction times in the lattice graph (right) relative to the modular graph (left). We detect a significant difference in reaction times for random walks ($p < 0.001$, $t = 3.95$, $df = 3.33 \times 10^5$); see Tab. 7.2 for the mixed effects model. Measurements were on independent subjects, statistical significance was computed using two-sided F-tests, and confidence intervals represent standard deviations. Source data are provided as a Source Data file.

(Figs. 6.8 and 6.9). Finally, for all of our analyses throughout the paper we regress out the dependence of reaction times on stimulus recency (see Methods). Together, these results demonstrate that higher-order network effects on human behavior cannot be explained by recency alone.

In combination, our experimental observations indicate that people are sensitive to the higher-order architecture of transition networks. But how do people infer abstract features like community structure from sequences of stimuli? In what follows, we turn to the free energy principle to show that a possible answer lies in understanding the subtle role of mental errors.

6.2.2 *Network effects reveal errors in graph learning*

As humans observe a sequence of stimuli or events, they construct an internal representation $\hat{A}$ of the transition structure, where $\hat{A}_{ij}$ represents the expected probability of transitioning from node i to node j. Given a running tally n_{ij} of the number of times each transition has occurred, one might naïvely expect that the human brain is optimized to learn the true transition structure as accurately as possible (*451*, *629*). This common hypothesis is represented by the maximum likelihood estimate (*93*), taking the simple form

$$\hat{A}_{ij}^{\mathrm{MLE}} = \frac{n_{ij}}{\sum_k n_{ik}}. \tag{6.1}$$

To see that human behavior does not reflect maximum likelihood estimation, we note that Eq. (6.1) provides an unbiased estimate of the transition structure (*93*); that is, the estimated transition probabilities in $\hat{A}^{\mathrm{MLE}}$ are evenly distributed about their true value 0.25, independent of the higher-order transition structure. Thus, the fact that people's reaction times depend systematically on abstract features of the network marks a clear deviation from maximum likelihood estimation. To understand how higher-order network structure impacts people's internal representations, we must delve deeper into the learning process itself.

Consider a sequence of nodes $(x_1, x_2, \ldots)$, where $x_t \in \{1, \ldots, N\}$ represents the node observed at time t and N is the size of the network (here $N = 15$ for all graphs). To update the maximum likelihood estimate of the transition structure at time $t+1$, one increments the counts n_{ij} using the following recursive rule,

$$n_{ij}(t+1) = n_{ij}(t) + [i = x_t]\,[j = x_{t+1}]\,, \tag{6.2}$$

where the Iverson bracket $[\cdot] = 1$ if its argument is true and 0 otherwise. Importantly, we note that at each time $t+1$, a person must recall the previous node that occurred at time t; in other words, they must associate a cause x_t to each effect x_{t+1} that they observe. While maximum likelihood estimation requires perfect recollection of the previous node at each step, human errors in perception and recall are inevitable (*277*, *323*, *324*). A more plausible scenario is that, when attempting to recall the node at time t, a person instead remembers the node at time $t - \Delta t$ with some decreasing probability $P(\Delta t)$, where $\Delta t \geqslant 0$. This memory distribution, in turn, generates an internal belief about which node occurred at time t,

$$B_t(i) = \sum_{\Delta t=0}^{t-1} P(\Delta t)\,[i = x_{t-\Delta t}]\,. \tag{6.3}$$

Updating Eq. (6.2) accordingly, we arrive at a learning rule that accounts for natural errors in perception and recall,

$$\tilde{n}_{ij}(t+1) = \tilde{n}_{ij}(t) + B_t(i)\,[j = x_{t+1}]\,. \tag{6.4}$$

Using this revised counting rule, we can begin to form more realistic predictions about people's internal estimates of the transition structure, $\hat{A}_{ij} = \tilde{n}_{ij} / \sum_k \tilde{n}_{ik}$.

We remark that $P(\Delta t)$ does not represent the forgetting of past stimuli altogether; instead, it reflects the local shuffling of stimuli in time. If one were to forget past stimuli at some fixed rate – a process that is important for some cognitive functions (*555*) – this would merely introduce white noise into the maximum likelihood estimate $\hat{A}^{MLE}$ (see Sec. 6.5.9). By contrast, we will see that, by shuffling the order of stimuli in time, people are able to gather information about the higher-order structure of the underlying transitions.

6.2.3 *Choosing a memory distribution: The free energy principle*

In order to make predictions about people's expectations, we must choose a particular mathematical form for the memory distribution $P(\Delta t)$. To do so, we begin with a single driving hypothesis: that the brain is finely-tuned to (i) minimize errors and (ii) minimize computational complexity. Formally, we define the error of a recalled stimulus to be its distance in time from the desired stimulus (i.e., Δt), such that the average error of a candidate distribution $Q(\Delta t)$ is given by $E(Q) = \sum_{\Delta t} Q(\Delta t)\Delta t$. By contrast, it might seem difficult to formalize the computational complexity associated with a distribution Q. Intuitively, we would like the complexity of Q to increase with increasing certainty. Moreover, as a first approximation we expect the complexity to be approximately additive such that the cost of storing two independent memories equals the costs of the two memories themselves. As famously shown by Shannon, these two criteria of monotonicity and additivity are sufficient to derive a quantitative definition of complexity (*603*) – namely, the negative entropy $-S(Q) = \sum_{\Delta t} Q(\Delta t) \log Q(\Delta t)$.

Together, the total cost of a distribution Q is its free energy $F(Q) = \beta E(Q) - S(Q)$, where β is the inverse temperature parameter, which quantifies the relative value that the brain places on accuracy versus efficiency (*493*). In this way, our assumption about resource constraints in the brain leads to a particular form for P: it should be the distribution that minimizes $F(Q)$, namely the Boltzmann distribution (*339*)

$$P(\Delta t) = \frac{1}{Z} e^{-\beta \Delta t}, \tag{6.5}$$

where Z is the normalizing constant (see Methods). Free energy arguments similar to the one presented here have been used increasingly to formalize constraints on cognitive functions (*232*, *493*), with applications from active inference (*233*) and Bayesian learning under uncertainty (*232*) to human action and perception with temporal or computational limitations (*248*, *493*, *494*). Taken together, Eqs. (6.3-6.5) define our maximum entropy model of people's internal transition estimates $\hat{A}$.

To gain an intuition for the model, we consider the infinite-time limit, such that the transition estimates become independent of the particular random walk chosen for analysis. Given a transition matrix A, one can show that the asymptotic estimates in our model are equivalent to an average over walks of various lengths,

$\hat{A} = \sum_{\Delta t} P(\Delta t) A^{\Delta t+1}$, which, in turn, can be fashioned into the following analytic expression,

$$\hat{A} = (1 - e^{-\beta}) A (I - e^{-\beta} A)^{-1}, \tag{6.6}$$

where I is the identity matrix (see Methods). The model contains a single free parameter β, which represents the precision of a person's mental representation. In the limit $\beta \to \infty$ (no mental errors), our model becomes equivalent to maximum likelihood estimation (Fig. 7.2a), and the asymptotic estimates $\hat{A}$ converge to the true transition structure A (Fig 7.2b), as expected (*281*). Conversely, in the limit $\beta \to 0$ (overwhelming mental errors), the memory distribution $P(\Delta t)$ becomes uniform across all past nodes (Fig. 7.2a), and the mental representation $\hat{A}$ loses all resemblance to the true structure A (Fig. 7.2b).

Remarkably, for intermediate values of β, higher-order features of the transition network, such as communities of densely-connected nodes, come into focus, while some of the fine-scale features, like the edges between communities, fade away (Fig. 7.2b). Applying Eq. (6.6) to the modular graph, we find that the average expected probability of within-community transitions reaches over 1.6 times the estimated probability of between-community transitions (Fig. 7.2c), thus explaining the cross-cluster surprisal effect (*351*, *361*). Furthermore, we find that the average estimated transition probabilities in the modular graph reach over 1.4 times the estimated probabilities in the lattice graph (Fig. 7.2d), thereby predicting the modular-lattice effect. In addition to these higher-order effects, we find that the model also explains previously reported variations in human expectations at the level of individual nodes (*217*, *351*, *576*) (Fig. 6.7). Together, these results demonstrate that the maximum entropy model predicts the qualitative effects of network structure on human reaction times. But can we use the same ideas to quantitatively predict the behavior of particular individuals?

6.2.4 *Predicting the behavior of individual humans*

To model the behavior of individual subjects, we relate the transition estimates in Eqs. (6.3-6.5) to predictions about people's reaction times. Given a sequence of nodes $x_1, \ldots, x_{t-1}$, we note that the reaction to the next node x_t is determined by the expected probability of transitioning from x_{t-1} to x_t calculated at time $t-1$, which we denote by $a(t) = \hat{A}_{x_{t-1}, x_t}(t-1)$. From this internal anticipation $a(t)$, the simplest possible prediction $\hat{r}(t)$ for a person's reaction time is given by the linear relationship (*472*) $\hat{r}(t) = r_0 + r_1 a(t)$, where the intercept r_0 represents a person's reaction time with zero anticipation and the slope r_1 quantifies the strength of the relationship between a person's reactions and their anticipation in our model (*597*).

To estimate the parameters β, r_0, and r_1 that best describe a given individual, we minimize the root mean squared error (RMSE) between their predicted and observed reaction times after regressing out the dependencies on button combination, trial number, and recency (Figs. 7.1e and 7.1f; see Methods). The distributions of the estimated parameters are shown in Fig. 6.4a-b for random walks and in Fig. 6.4g-h

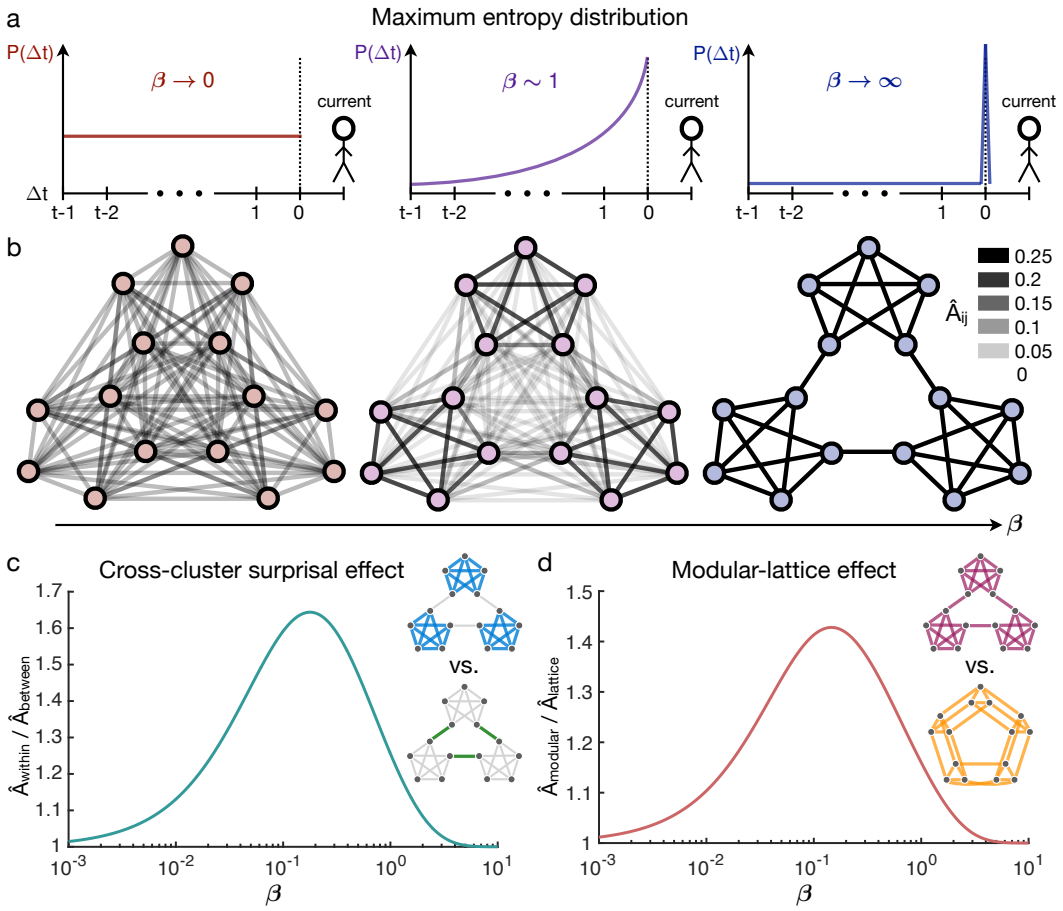


Figure 6.3: A maximum entropy model of transition probability estimates in humans. (*a*) Illustration of the maximum entropy distribution $P(\Delta t)$ representing the probability of recalling a stimulus Δt time steps from the target stimulus (dashed line). In the limit $\beta \to 0$, the distribution becomes uniform over all past stimuli (left). In the opposite limit $\beta \to \infty$, the distribution becomes a delta function on the desired stimulus (right). For intermediate amounts of noise, the distribution drops off monotonically (center). (*b*) Resulting internal estimates $\hat{A}$ of the transition structure. For $\beta \to 0$, the estimates become all-to-all, losing any resemblance to the true structure (left), while for $\beta \to \infty$, the transition estimates become exact (right). At intermediate precision, the higher-order community structure organically comes into focus (center). (*c-d*) Predictions of the cross-cluster surprisal effect (*c*) and the modular-lattice effect (*d*) as functions of the inverse temperature β.

for Hamiltonian walks. Among the 358 random walk sequences in the modular and lattice graphs (across 286 subjects; see Methods), 40 were best described as performing maximum likelihood estimation ($\beta \to \infty$) and 73 seemed to lack any notion of the transition structure whatsoever ($\beta \to 0$), while among the remaining 245 sequences, the average inverse temperature was $\beta = 0.30$. Meanwhile, among the 120 subjects that responded to Hamiltonian walk sequences, 81 appeared to have a non-trivial value of β, with an average of $\beta = 0.61$. Interestingly, these estimates of β roughly correspond to the values for which our model predicts the strongest network effects

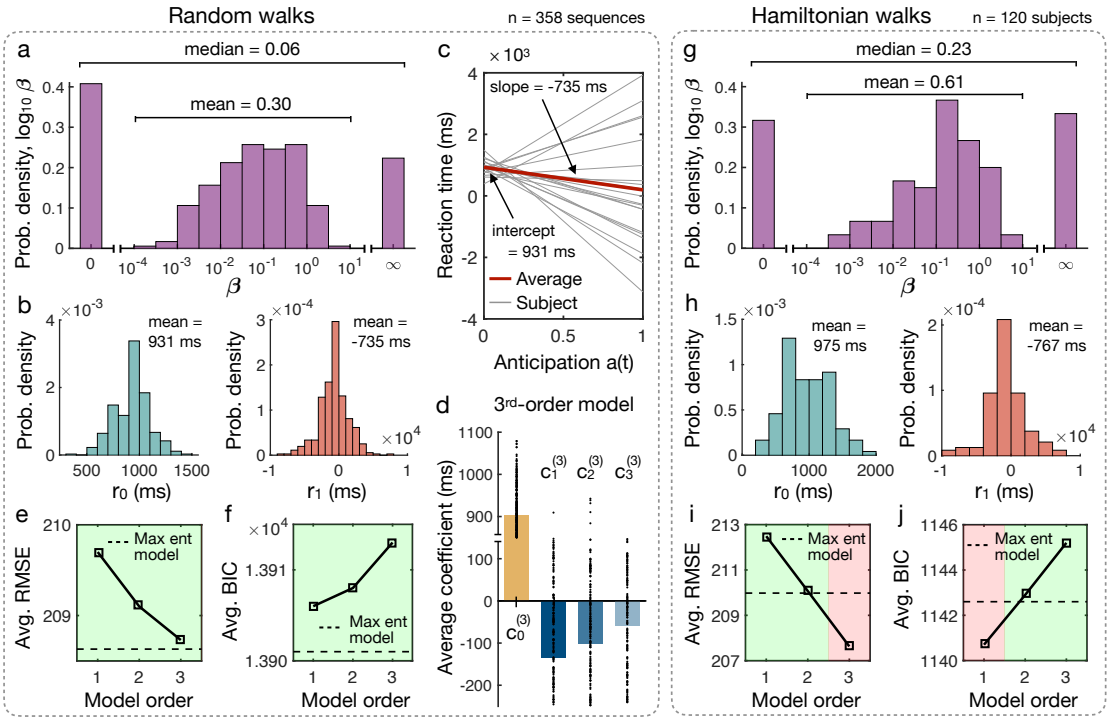


Figure 6.4: Predicting reaction times for individual subjects. (*a-f*) Estimated parameters and accuracy analysis for our maximum entropy model across 358 random walk sequences (across 286 subjects; see Methods). (*a*) For the inverse temperature β, 40 sequences corresponded to the limit $\beta \to \infty$, 73 corresponded to the limit $\beta \to 0$. Among the remaining 245 sequences, the average value of β was 0.30. (*b*) Distributions of the intercept r_0 (left) and slope r_1 (right). (*c*) Predicted reaction time as a function of a subject's internal anticipation. Grey lines indicate 20 randomly-selected sequences, and the red line shows the average prediction over all sequences. (*d*) Linear parameters for the third-order competing model; data points represent individual sequences and bars represent averages. (*e-f*) Comparing the performance of our maximum entropy model with the hierarchy of competing models up to third-order. Root mean squared error (RMSE; *e*) and Bayesian information criterion (BIC; *f*) of our model averaged over all sequences (dashed lines) compared to the competing models (solid lines); our model provides the best description of the data across all models considered. (*g-j*) Estimated parameters and accuracy analysis for our maximum entropy model across all Hamiltonian walk sequences (120 subjects). (*g*) For the inverse temperature β, 20 subjects were best described as performing maximum likelihood estimation ($\beta \to \infty$), 19 lacked any notion of the transition structure ($\beta \to 0$), and the remaining 81 subjects had an average value of $\beta = 0.61$. (*h*) Distributions of the intercept r_0 (left) and slope r_1 (right). (*i*) Average RMSE of our model (dashed line) compared to that of the competing models (solid line); our model maintains higher accuracy than the competing hierarchy up to the second-order model. (*j*) Average BIC of the maximum entropy model (dashed line) compared to that of the competing models (solid line); our model provides a better description of the data than the second- or third-order models. Source data are provided as a Source Data file.

(Figs. 7.2c and 7.2d). In the following section, we will compare these values of β, which are estimated indirectly from people's reaction times, with direct measurements of β in an independent memory experiment.

In addition to estimating β, we also wish to determine whether our model accurately describes individual behavior. Toward this end, we first note that the average slope r_1 is large (-735 ms for random walks and -767 ms for Hamiltonian walks), suggesting that the transition estimates in our model $a(t)$ are strongly predictive of human reaction times, and negative, confirming the intuition that increased anticipation yields decreased reaction times (Figs. 6.4b and 6.4h). To examine the accuracy of our model $\hat{r}$, we consider a hierarchy of competing models $\hat{r}^{(\ell)}$, which represent the hypothesis that humans learn explicit representations of the higher-order transition structure. In particular, we denote the ℓ^{th}-order transition matrix by $\hat{A}^{(\ell)}_{ij} = n^{(\ell)}_{ij} / \sum_k n^{(\ell)}_{ik}$, where $n^{(\ell)}_{ij}$ counts the number of observed transitions from node i to node j in ℓ steps. The model hierarchy takes into account increasingly higher-order transitions, such that the ℓ^{th}-order model contains perfect information about transitions up to length ℓ:

$$\begin{aligned}
\hat{r}^{(0)}(t) &= c^{(0)}_0, \\
\hat{r}^{(1)}(t) &= c^{(1)}_0 + c^{(1)}_1 a^{(1)}(t), \\
&\vdots \\
\hat{r}^{(\ell)}(t) &= c^{(\ell)}_0 + \sum_{k=1}^{\ell} c^{(\ell)}_k a^{(k)}(t),
\end{aligned} \tag{6.7}$$

where $a^{(k)}(t) = \hat{A}^{(k)}_{x_{t-1}, x_t}(t-1)$. Each model $\hat{r}^{(\ell)}$ contains $\ell + 1$ parameters $c^{(\ell)}_0, \ldots, c^{(\ell)}_\ell$, where $c^{(\ell)}_k$ quantifies the predictive power of the k^{th}-order transition structure.

Intuitively, for each model $\hat{r}^{(\ell)}$, we expect $c^{(\ell)}_1, c^{(\ell)}_2, \ldots$ to be negative, reflecting a decrease in reaction times due to increased anticipation, and decreasing in magnitude, such that higher-order transitions are progressively less predictive of people's reaction times. Indeed, considering the third-order model $\hat{r}^{(3)}$ as an example, we find that progressively higher-order transitions are less predictive of human reactions (Fig. 6.4d). However, even the largest coefficient ($c^{(3)}_1 = -135$ ms) is much smaller than the slope in our maximum entropy model ($r_1 = -735$ ms), indicating that the representation $\hat{A}$ is more strongly predictive of people's reaction times than any of the explicit representations $\hat{A}^{(1)}, \hat{A}^{(2)}, \ldots$. Indeed, averaging over the random walk sequences, the maximum entropy model achieves higher accuracy than the first three orders of the competing model hierarchy (Fig. 6.4e) – this is despite the fact that the third-order model even contains one more parameter. To account for differences in the number of parameters, we additionally compare the average Bayesian information criterion (BIC) of our model with that of the competing models, finding that the maximum entropy model provides the best description of the data (Fig. 6.4f).

Similarly, averaging over the Hamiltonian walk sequences, the maximum entropy model provides more accurate predictions than the first two competing models (Fig. 6.4i) and provides a lower BIC than the second and third competing models (Fig. 6.4j). Notably, even in Hamiltonian walks, the maximum entropy model provides a better

description of human reaction times than the second-order competing model, which has the same number of parameters. However, we remark that the first-order competing model has a lower BIC than the maximum entropy model (Fig. 6.4j), suggesting that humans may focus on first-order rather than higher-order statistics during Hamiltonian walks – an interesting direction for future research. On the whole, these results indicate that the free energy principle, and the resulting maximum entropy model, are consistently more effective at describing human reactions than the hypothesis that people learn explicit representations of the higher-order transition structure.

6.2.5 *Directly probing the memory distribution*

Throughout our discussion, we have argued that errors in memory shape human representations in predictable ways, a perspective that has received increasing attention in recent years (*153*, *154*, *555*). While our framework explains specific aspects of human behavior, there exist alternative perspectives that might yield similar predictions. For example, one could imagine a Bayesian learner with a non-Markov prior that "integrates" the transition structure over time, even without sustaining errors in memory or learning. Additionally, Eq. (6.6) resembles the successor representation in reinforcement learning (*170*, *247*), which assumes that, rather than shuffling the order of past stimuli, humans are instead planning their responses multiple steps in advance (see Sec. 6.5.11). In order to distinguish our framework from these alternatives, here we provide direct evidence for precisely the types of mental errors predicted by our model.

In the construction and testing of our model, we have developed a series of predictions concerning the shape of the memory distribution $P(\Delta t)$, which, to recall, represents the probability of remembering the stimulus at time $t - \Delta t$ instead of the target stimulus at time t. We first assumed that $P(\Delta t)$ decreases monotonically. Second, to make quantitative predictions, we employed the free energy principle, leading to the prediction that P drops off exponentially quickly with Δt (Eq. (6.5)). Finally, when fitting the model to individual subjects, we estimated an average inverse temperature β between 0.30 for random walks and 0.61 for Hamiltonian walks.

To test these three predictions directly, we conducted a standard n-back memory experiment. Specifically, we presented subjects with sequences of letters on a screen, and they were asked to respond to each letter indicating whether or not it was the same as the letter that occurred n steps previously; for each subject, this process was repeated for the three conditions $n = 1, 2$, and 3. To measure the memory distribution $P(\Delta t)$, we considered all trials on which a subject responded positively that the current stimulus matched the target. For each such trial, we looked back to the last time that the subject did in fact observe the current stimulus and we recorded the distance (in trials) between this observation and the target (Fig. 6.5a). In this way, we were able to treat each positive response as a sample from the memory distribution $P(\Delta t)$.

The measurements of P for the 1-, 2-, and 3-back tasks are shown in Figure 6.5b, and the combined measurement of P across all conditions is shown in Figure 6.5c. Notably, the distributions decrease monotonically and maintain consistent exponential forms,

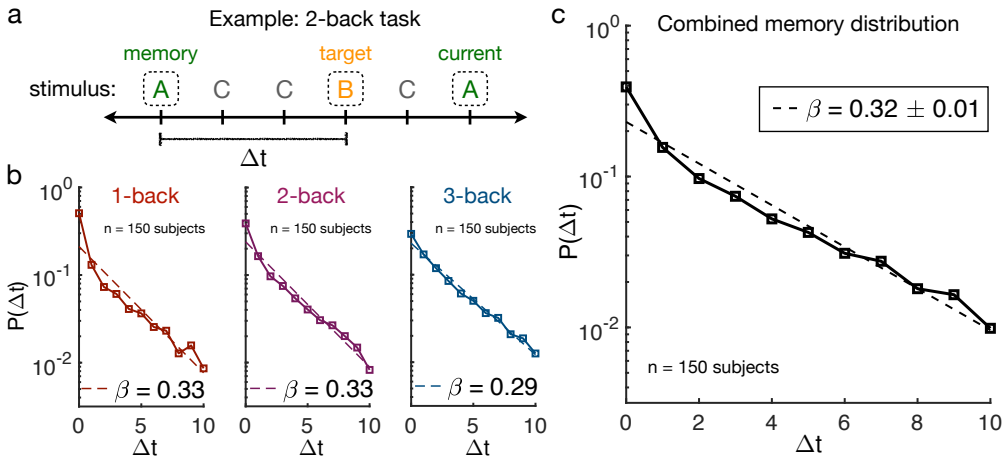


Figure 6.5: Measuring the memory distribution in an n-back experiment. (*a*) Example of the 2-back memory task. Subjects view a sequence of stimuli (letters) and respond to each stimulus indicating whether it matches the target stimulus from two trials before. For each positive response that the current stimulus matches the target, we measure Δt by calculating the number of trials between the last instance of the current stimulus and the target. (*b*) Histograms of Δt (i.e., measurements of the memory distribution $P(\Delta t)$) across all subjects in the 1-, 2-, and 3-back tasks. Dashed lines indicate exponential fits to the observed distributions. The inverse temperature β is estimated for each task to be the negative slope of the exponential fit. (*c*) Memory distribution aggregated across the three n-back tasks. Dashed line indicates an exponential fit. We report a combined estimate of the inverse temperature $\beta = 0.32 \pm 0.01$, where the standard deviation is estimated from 1,000 bootstrap samples of the combined data. Measurements were on independent subjects. Source data are provided as a Source Data file.

even out to $\Delta t = 10$ trials from the target stimulus, thereby providing direct evidence for the Boltzmann distribution (Eq. (6.5)). Moreover, fitting an exponential curve to each distribution, we can directly estimate the inverse temperature β. Remarkably, the value $\beta = 0.32 \pm 0.1$ estimated from the combined distribution (Fig. 6.5c) falls within the range of values estimated from our reaction time experiments (Figs. 6.4a and 6.4g), nearly matching the average value $\beta = 0.30$ for random walk sequences (Fig. 6.4a).

To further strengthen the link between mental errors and people's internal representations, we then asked subjects to perform the original serial response task (Fig. 7.1), and for each subject, we estimated β using the two methods described above: (i) directly measuring β in the n-back experiment, and (ii) indirectly estimating β in the serial response experiment. Comparing these two estimates across subjects, we find that they are significantly related with Spearman correlation $r_s = 0.28$ ($p = 0.047$, permutation test), while noting that we do not use the Pearson correlation because β is not normally distributed (Anderson-Darling test (*634*), $p < 0.001$ for the serial response task and $p = 0.013$ for the n-back task). Together, these results demonstrate not only the existence of the particular form of mental errors predicted by our model – down to the specific value of β – but also the relationship between these mental errors and people's internal estimates of the transition structure.

6.2.6 *Network structure guides reactions to novel transitions*

Given a model of human behavior, it is ultimately interesting to make testable predictions. Thus far, in keeping with the majority of existing research (*217*, *351*, *360*, *361*, *576*, *584*), we have focused on static transition graphs, wherein the probability A_{ij} of transitioning from state i to state j remains constant over time. However, the statistical structures governing human life are continually shifting (*696*, *706*), and people are often forced to respond to rare or novel transitions (*672*, *712*). Here we show that, when confronted with a novel transition – or a *violation* of the preexisting transition network – not only are people surprised, but the magnitude of their surprise depends critically on the topology of the underlying network.

We consider a ring graph where each node is connected to its nearest and next-nearest neighbors (Fig. 6.5.7a). We asked subjects to respond to sequences of 1500 nodes drawn as random walks on the ring graph, but with 50 violations randomly interspersed. These violations were divided into two categories: short violations of topological distance two and long violations of topological distances three and four (Fig. 6.5.7a). Using maximum likelihood estimation (Eq. (6.1)) as a guide, one would naïvely expect people to be equally surprised by all violations – indeed, each violation has never been seen before. In contrast, our model predicts that that surprise should depend crucially on the topological distance of a violation in the underlying graph, with topologically longer violations inducing increased surprise over short violations (Fig. 6.5.7b).

In the data, we find that all violations give rise to sharp increases in reaction times relative to standard transitions (Fig. 6.5.7c; Tab. 7.10), indicating that people are in fact learning the underlying transition structure. Moreover, we find that reaction times for long violations are 28 ms longer than those for short violations ($p = 0.011$, F-test; Fig. 6.5.7c; Tab. 7.11). Additionally, we confirm that the effects of network violations are not simply driven by stimulus recency (Figs. 6.10 and 6.11). These observations suggest that people learn the topological distances between all nodes in the transition graph, not just those pairs for which a transition has already been observed (*672*, *696*, *706*, *712*).

6.3 DISCUSSION

Daily life is filled with sequences of items that obey an underlying network architecture, from networks of word and note transitions in natural language and music to networks of abstract relationships in classroom lectures and literature (*99*, *217*, *227*, *554*, *667*). How humans infer and internally represent these complex structures are questions of fundamental interest (*180*, *444*, *524*, *658*). Recent experiments in statistical learning have established that human representations depend critically on the higher-order organization of probabilistic transitions, yet the underlying mechanisms remain poorly understood (*351*, *360*, *362*, *584*).

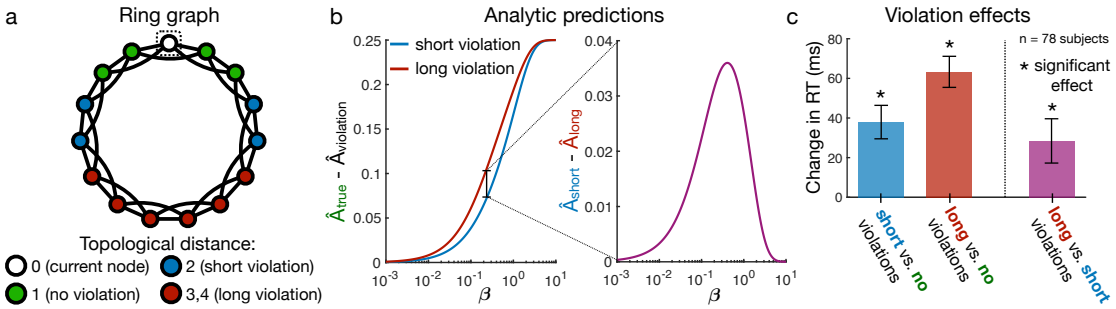


Figure 6.6: Network violations yield surprise that grows with topological distance. (*a*) Ring graph consisting of 15 nodes, where each node is connected to its nearest neighbors and next-nearest neighbors on the ring. Starting from the boxed node, a sequence can undergo a standard transition (green), a short violation of the transition structure (blue), or a long violation (red). (*b*) Our model predicts that subjects' anticipations of both short (blue) and long (red) violations should be weaker than their anticipations of standard transitions (left). Furthermore, we predict that subjects' anticipations of violations should decrease with increasing topological distance (right). (*c*) Average effects of network violations across 78 subjects, estimated using a mixed effects model (see Tabs. 7.10 and 7.11), with error bars indicating one standard deviation from the mean. We find that standard transitions yield quicker reactions than both short violations ($p < 0.001$, $t = 4.50$, $df = 7.15 \times 10^4$) and long violations ($p < 0.001$, $t = 8.07$, $df = 7.15 \times 10^4$). Moreover, topologically shorter violations induce faster reactions than long violations ($p = 0.011$, $t = 2.54$, $df = 3.44 \times 10^3$), thus confirming the predictions of our model. Measurements were on independent subjects, and statistical significance was computed using two-sided F-tests. Source data are provided as a Source Data file.

Here we show that network effects on human behavior can be understood as stemming from mental errors in people's estimates of the transition structure, while noting that future work should focus on disambiguating the role of recency (*41*, *465*). We use the free energy principle to develop a model of human expectations that explicitly accounts for the brain's natural tendency to minimize computational complexity – that is, to maximize entropy (*232*, *493*, *494*). Indeed, the brain must balance the benefits of making accurate predictions against the computational costs associated with such predictions (*150*, *174*, *248*, *669*, *674*, *686*, *708*). This competition between accuracy and efficiency induces errors in people's internal representations, which, in turn, explains with notable accuracy an array of higher-order network phenomena observed in human experiments (*351*, *360*, *362*, *584*). Importantly, our model admits a concise analytic form (Eq. (6.6)) and can be used to predict human behavior on a person-by-person basis (Fig. 6.4).

This work inspires directions for future research, particularly with regard to the study and design of optimally learnable network structures. Given the notion that densely connected communities help to mitigate the effects of mental errors on people's internal representations, we anticipate that networks with high "learnability" will possess a hierarchical community structure (*25*). Interestingly, such hierarchical organization has already been observed in a diverse range of real world networks, from knowledge and language graphs (*286*) to social networks and the World Wide Web (*547*). Could it

be that these networks have evolved so as to facilitate accurate representations in the minds of the humans using and observing them? Questions such as this demonstrate the importance of having simple principled models of human representations and point to the promising future of this research endeavor.

6.4 METHODS

6.4.1 *Maximum entropy model and the infinite-sequence limit*

Here we provide a more thorough derivation of our maximum entropy model of human expectations, with the goal of fostering intuition. Given a matrix of erroneous transition counts $\tilde{n}_{ij}$, our estimate of the transition structure is given by $\hat{A}_{ij} = \tilde{n}_{ij} / \sum_k \tilde{n}_{ik}$. When observing a sequence of nodes $x_1, x_2, \ldots$, in order to construct the counts $\tilde{n}_{ij}$, we assume that humans use the following recursive rule: $\tilde{n}_{ij}(t+1) = \tilde{n}_{ij}(t) + B_t(i)\,[j = x_{t+1}]$, where $B_t(i)$ denotes the belief, or perceived probability, that node i occurred at the previous time t. This belief, in turn, can be written in terms of the probability $P(\Delta t)$ of accidentally recalling the node that occurred Δt time steps from the desired node at time t: $B_t(i) = \sum_{\Delta t=0}^{t-1} P(\Delta t)\,[i = x_{t-\Delta t}]$.

In order to make quantitative predictions about people's estimates of a transition structure, we must choose a mathematical form for $P(\Delta t)$. To do so, we leverage the free energy principle (*493*): When building mental models, the brain is finely-tuned to simultaneously minimize errors and computational complexity. The average error associated with a candidate distribution $Q(\Delta t)$ is assumed to be the average distance in time of the recalled node from the target node, denoted $E(Q) = \sum_{\Delta t} Q(\Delta t)\Delta t$. Furthermore, Shannon famously proved that the only suitable choice for the computational cost of a candidate distribution is its negative entropy (*603*), denoted $-S(Q) = \sum_{\Delta t} Q(\Delta t) \log Q(\Delta t)$. Taken together, the total cost associated with a distribution $Q(\Delta t)$ is given by the free energy $F(Q) = \beta E(Q) - S(Q)$, where β, referred to as the inverse temperature, parameterizes the relative importance of minimizing errors versus computational costs. By minimizing F with respect to Q, we arrive at the Boltzmann distribution $P(\Delta t) = e^{-\beta \Delta t}/Z$, where Z is the normalizing partition function (*339*). We emphasize that this mathematical form for $P(\Delta t)$ followed directly from our free energy assumption about resource constraints in the brain.

To gain an analytic intuition for the model without referring to a particular random walk, we consider the limit of an infinitely long sequence of nodes. To begin, we consider a sequence $x_1, \ldots, x_T$ of length T. At the end of this sequence, the counting matrix takes the form

$$\begin{aligned} \tilde{n}_{ij}(T) &= \sum_{t=1}^{T-1} B_t(i)\,[j = x_{t+1}] \\ &= \sum_{t=1}^{T-1} \left(\sum_{\Delta t=0}^{t-1} P(\Delta t)\,[i = x_{t-\Delta t}] \right) [j = x_{t+1}]\,. \end{aligned} \tag{6.8}$$

Dividing both sides by T, the right-hand side becomes a time average, which by the ergodic theorem converges to an expectation over the transition structure in the limit $T \to \infty$,

$$\lim_{T\to\infty} \frac{\tilde{n}_{ij}(T)}{T} = \sum_{\Delta t=0}^{\infty} P(\Delta t) \left\langle [i = x_{t-\Delta t}]\, [j = x_{t+1}] \right\rangle_A , \tag{6.9}$$

where $\langle\cdot\rangle_A$ denotes an expectation over random walks in A. We note that the expectation of an identity function is simply a probability, such that $\left\langle [i = x_{t-\Delta t}]\, [j = x_{t+1}] \right\rangle_A = p_i \left(A^{\Delta t+1}\right)_{ij}$, where p_i is the long-run probability of node i appearing in the sequence and $\left(A^{\Delta t+1}\right)_{ij}$ is the probability of randomly walking from node i to node j in $\Delta t + 1$ steps. Putting these pieces together, we find that the expectation $\hat{A}$ converges to a concise mathematical form,

$$\begin{aligned}
\lim_{T\to\infty} \hat{A}_{ij}(T) &= \lim_{T\to\infty} \frac{\tilde{n}_{ij}(T)}{\sum_k \tilde{n}_{ik}(T)} \\
&= \frac{p_i \sum_{\Delta t=0}^{\infty} P(\Delta t) \left(A^{\Delta t+1}\right)_{ij}}{p_i} \\
&= \sum_{\Delta t=0}^{\infty} P(\Delta t) \left(A^{\Delta t+1}\right)_{ij} .
\end{aligned} \tag{6.10}$$

Thus far, we have not appealed to our maximum entropy form for $P(\Delta t)$. It turns out that doing so allows us to write down an analytic expression for the long-time expectations $\hat{A}$ simply in terms of the transition structure A and the inverse temperature β. Noting that $Z = \sum_{\Delta t=0}^{\infty} e^{-\beta \Delta t} = 1/(1 - e^{-\beta})$ and $\sum_{\Delta t=0}^{\infty} \left(e^{-\beta} A\right)^{\Delta t} = \left(I - e^{-\beta} A\right)^{-1}$, we have

$$\begin{aligned}
\hat{A} &= \sum_{\Delta t=0}^{\infty} P(\Delta t) A^{\Delta t+1} \\
&= \frac{1}{Z} A \sum_{\Delta t=0}^{\infty} \left(e^{-\beta} A\right)^{\Delta t} \\
&= \left(1 - e^{-\beta}\right) A \left(I - e^{-\beta} A\right)^{-1} .
\end{aligned} \tag{6.11}$$

This simple formula for the representation $\hat{A}$ is the basis for all of our analytic predictions (Figs. 7.2c, 7.2d, and 6.5.7b) and is closely related to notions of communicability in complex network theory (*209*, *210*).

6.4.2 *Experimental setup for serial response tasks*

Subjects performed a self-paced serial reaction time task using a computer screen and keyboard. Each stimulus was presented as a horizontal row of five grey squares; all five squares were shown at all times. The squares corresponded spatially with the keys 'Space', 'H', 'J', 'K', and 'L', with the left square representing 'Space' and the

right square representing 'L' (Fig. 7.1b). To indicate a target key or pair of keys for the subject to press, the corresponding squares would become outlined in red (Fig. 7.1a). When subjects pressed the correct key combination, the squares on the screen would immediately display the next target. If an incorrect key or pair of keys was pressed, the message 'Error!' was displayed on the screen below the stimuli and remained until the subject pressed the correct key(s). The order in which stimuli were presented to each subject was prescribed by either a random walk or a Hamiltonian walk on a graph of $N = 15$ nodes, and each sequence consisted of 1500 stimuli. For each subject, one of the 15 key combinations was randomly assigned to each node in the graph (Fig. 7.1a). Across all graphs, each node was connected to its four neighboring nodes with a uniform 0.25 transition probability. Importantly, given the uniform edge weights and homogeneous node degrees ($k = 4$), the only differences between the transition graphs lay in their higher-order structure.

In the first experiment, we presented subjects with random walk sequences drawn from two different graph topologies: a *modular* graph with three communities of five densely-connected nodes and a *lattice* graph representing a 3×5 grid with periodic boundary conditions (Fig. 7.1c). The purpose of this experiment was to demonstrate the systematic dependencies of human reaction times on higher-order network structure, following similar results reported in recent literature (*351*, *361*). In particular, we demonstrate two higher-order network effects: In the *cross-cluster surprisal* effect, average reaction times for within-cluster transitions in the modular graph are significantly faster than reaction times for between-cluster transitions (Fig. 6.2a); and in the *modular-lattice* effect, average reaction times in the modular graph are significantly faster than reaction times in the lattice graph (Fig. 6.2b).

In the second experiment, we presented subjects with Hamiltonian walk sequences drawn from the modular graph. Specifically, each sequence consisted of 700 random walk trials (intended to allow each subject to learn the graph structure), followed by eight repeats of 85 random walk trials and 15 Hamiltonian walk trials (*584*). Importantly, we find that the cross-cluster surprisal effect remains significant within the Hamiltonian walk trials (Fig. 6.2a).

In the third experiment, we considered a *ring* graph where each node was connected to its nearest and next-nearest neighbors in the ring (Fig. 6.5.7a). In order to study the dependence of human expectations on violations to the network structure, the first 500 trials for each subject constituted a standard random walk, allowing each subject time to develop expectations about the underlying transition structure. Across the final 1000 trials, we randomly distributed 50 network violations: 20 short violations of topological distance two and 30 long violations, 20 of topological distance three and 10 of topological distance four (Fig. 6.5.7a). As predicted by our model, we found a novel *violations* effect, wherein violations of longer topological distance give rise to larger increases in reaction times than short, local violations (Figs. 6.5.7b and 6.5.7c).

6.4.3 *Data analysis for serial response tasks*

To make inferences about subjects' internal expectations based on their reaction times, we used more stringent filtering techniques than previous experiments when pre-processing the data (*351*). Across all experiments, we first excluded from analysis the first 500 trials, in which subjects' reaction times varied wildly (Fig. 7.1e), focusing instead on the final 1000 trials (or simply on the Hamiltonian trials in the second experiment), at which point subjects had already developed internal expectations about the transition structures. We then excluded all trials in which subjects responded incorrectly. Finally, we excluded reaction times that were implausible, either three standard deviations from a subject's mean reaction time or below 100 ms. Furthermore, when measuring the network effects in all three experiments (Figs. 6.2 and 6.5.7), we also excluded reaction times over 3500 ms for implausibility. When estimating the parameters of our model and measuring model performance in the first two experiments (Fig. 6.4), to avoid large fluctuations in the results based on outlier reactions, we were even more stringent, excluding all reaction times over 2000 ms. Taken together, when measuring the cross-cluster surprisal and modular-lattice effects (Fig. 6.2), we used an average of 931 trials per subject; when estimating and evaluating our model (Fig. 6.4), we used an average of 911 trials per subject; and when measuring the violation effects (Fig. 6.5.7), we used an average of 917 trials per subject. To ensure that our results are robust to particular choices in the data processing, we additionally studied all 1500 trials for each subject rather than just the final 1000, confirming that both the cross-cluster surprisal and modular-lattice effects remain significant across all trials (Tabs. 7.6 and 7.7).

6.4.4 *Measurement of network effects using mixed effects models*

In order to extract the effects of higher-order network structure on subjects' reaction times, we used linear mixed effects models, which have become prominent in human research where many measurements are made for each subject (*39*, *582*). Put simply, mixed effects models generalize standard linear regression techniques to include both *fixed* effects, which are constant across subjects, and *random* effects, which vary between subjects. Compared with standard linear models, mixed effects models allow for differentiation between effects that are subject-specific and those that persist across an entire population. Here, all models were fit using the `fitlme` function in MATLAB (R2018a), and random effects were chosen as the maximal structure that (i) allowed model convergence and (ii) did not include effects whose 95% confidence intervals overlapped with zero (*326*). In what follows, when defining mixed effects models, we employ the standard R notation (*65*).

First, we considered the cross-cluster surprisal effect (Fig. 6.2a). Since we were only interested in measuring higher-order effects of the network topology on human reaction times, it was important to regress out simple biomechanical dependencies on the target button combinations (Fig. 7.1d), the natural quickening of reactions with

time (Fig. 7.1e), and the effects of recency on reaction times (*41*, *465*). Also, for the first experiment, since some subjects responded to both the modular and lattice graphs (see Experimental Procedures), it was important to account for changes in reaction times due to which stage of the experiment a subject was in. To measure the cross-cluster surprisal effect, we fit a mixed effects model with the formula 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Trans\_Type} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} + \text{Trans\_Type} | \text{ID})$', where RT is the reaction time, Trial is the trial number (we found that $\log(\text{Trial})$ was far more predictive of subjects' reaction times than the trial number itself), Stage is the stage of the experiment (either one or two), Target is the target button combination, Recency is the number of trials since the last instance of the current stimulus, Trans_Type is the type of transition (either within-cluster or between-cluster), and ID is each subject's unique ID. Fitting this mixed effects model to the random walk data in the first experiment (Tab. 7.1), we found a 35 ms increase in reaction times ($p < 0.001$, F-test) for between-cluster transitions relative to within-cluster transitions (Fig. 6.2a). Similarly, fitting the same mixed effects model but without the variable $Stage$ to the Hamiltonian walk data in the second experiment (Tab. 7.4), we found a 36 ms increase in reaction times ($p < 0.001$, F-test) for between- versus within-cluster transitions (Fig. 6.2a). We note that because reaction times are not Gaussian distributed, it is fairly standard to perform a log transformation. However, for the above result as well as those that follow, we find the same qualitative effects with or without a log transformation.

Second, we studied the modular-lattice effect (Fig. 6.2b). To do so, we fit a mixed effects model with the formula 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Graph} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} + \text{Graph} | \text{ID})$', where Graph represents the type of transition network, either modular or lattice. Fitting this mixed effects model to the data in the first experiment (Tab. 7.2), we found a fixed 23 ms increase in reaction times ($p < 0.001$, F-test) in the lattice graph relative to the modular graph (Fig. 6.2b).

Finally, we considered the effects of violations of varying topological distance in the ring lattice (Fig. 6.5.7c). We fit a mixed effects model with the formula 'RT $\sim \log(\text{Trial}) + \text{Target} + \text{Recency} + \text{Top\_Dist} + (1 + \log(\text{Trial}) + \text{Recency} + \text{Top\_Dist} | \text{ID})$', where Top_Dist represents the topological distance of a transition, either one for a standard transition, two for a short violation, or three for a long violation. Fitting the model to the data in the third experiment (Tabs. 7.10 and 7.11), we found a 38 ms increase in reaction times for short violations relative to standard transitions ($p < 0.001$, F-test), a 63 ms increase in reaction times for long violations relative to standard transitions ($p < 0.001$, F-test), and a 28 ms increase in reaction times for long violations relative to short violations ($p = 0.011$, F-test). Put simply, people are more surprised by violations to the network structure that take them further from their current position in the network, suggesting that people have an implicit understanding of the topological distances between nodes in the network.

6.4.5 *Estimating parameters and making quantitative predictions*

Given an observed sequence of nodes $x_1, \ldots, x_{t-1}$, and given an inverse temperature β, our model predicts the anticipation, or expectation, of the subsequent node x_t to be $a(t) = \hat{A}_{x_{t-1}, x_t}(t-1)$. In order to quantitatively describe the reactions of an individual subject, we must relate the expectations $a(t)$ to predictions about a person's reaction times $\hat{r}(t)$ and then calculate the model parameters that best fit the reactions of an individual subject. The simplest possible prediction is given by the linear relation $\hat{r}(t) = r_0 + r_1 a(t)$, where the intercept r_0 represents a person's reaction time with zero anticipation and the slope r_1 quantifies the strength with which a person's reaction times depend on their internal expectations.

In total, our predictions $\hat{r}(t)$ contain three parameters (β, r_0, and r_1), which must be estimated from the reaction time data for each subject. Before estimating these parameters, however, we first regress out the dependencies of each subject's reaction times on the button combinations, trial number, and recency using a mixed effects model of the form 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \,|\, \text{ID})$', where all variables were defined in the previous section. Then, to estimate the model parameters that best describe an individual's reactions, we minimize the RMS prediction error with respect to each subject's observed reaction times, $\text{RMSE} = \sqrt{\frac{1}{T}\sum_t (r(t) - \hat{r}(t))^2}$, where T is the number of trials. We note that, given a choice for the inverse temperature β, the linear parameters r_0 and r_1 can be calculated analytically using standard linear regression techniques. Thus, the problem of estimating the model parameters can be restated as a one-dimensional minimization problem; that is, minimizing RMSE with respect to the inverse temperature β. To find the global minimum, we began by calculating RMSE along 100 logarithmically-spaced values for β between 10^{-4} and 10. Then, starting at the minimum value of this search, we performed gradient descent until the gradient fell below an absolute value of 10^{-6}. For a derivation of the gradient of the RMSE with respect to the inverse temperature β, we point the reader to Sec. 6.5.10. Finally, in addition to the gradient descent procedure described above, for each subject we also manually checked the RMSE associated with the two limits $\beta \to 0$ and $\beta \to \infty$. The resulting model parameters are shown in Figs. 6.4a and 6.4b for random walk sequences and Figs. 6.4g and 6.4h for Hamiltonian walk sequences.

6.4.6 *Experimental setup for n-back memory task*

Subjects performed a series of n-back memory tasks using a computer screen and keyboard. Each subject observed a random sequence of the letters 'B', 'D', 'G', 'T', and 'V', wherein each letter was randomly displayed in either upper or lower case. The subjects responded on each trial using the keyboard to indicate whether or not the current letter was the same as the letter that occurred n trials previously. For each subject, this task was repeated for the conditions $n = 1, 2$, and 3, and each condition

consisted of a sequence of 100 letters. The three conditions were presented in a random order to each subject. After the n-back task, each subject then performed a serial response task (equivalent to the first experiment described above) consisting of 1500 random walk trials drawn from the modular graph.

6.4.7 *Data analysis for n-back memory task*

In order to estimate the inverse temperature β for each subject from their n-back data, we directly measured their memory distribution $P(\Delta t)$. As described in the main text, we treated each positive response indicating that the current stimulus matched the target stimulus as a sample of $P(\Delta t)$ by measuring the distance in trials Δt between the last instance of the current stimulus and the target (Fig. 6.5a). For each subject, we combined all such samples across the three conditions $n = 1, 2$, and 3 to arrive at a histogram for Δt. In order to generate robust estimates for the inverse temperature β, we generated 1000 bootstrap samples of the Δt histogram for each subject. For each sample, we calculated a linear fit to the distribution $P(\Delta t)$ on log-linear axes within the domain $0 \leqslant \Delta t \leqslant 4$ (note that we could not carry the fit out to $\Delta t = 10$ because the data is much sparser for individual subjects). To ensure that the logarithm of $P(\Delta t)$ was well defined for each sample – that is, to ensure that $P(\Delta t) > 0$ for all Δt – we added one count to each value of Δt. We then estimated the inverse temperature β for each sample by calculating the negative slope of the linear fit of $\log P(\Delta t)$ versus Δt. To arrive at an average estimate of β for each subject, we averaged β across the 1000 bootstrap samples. Finally, we compared these estimates of β from the n-back experiment with estimates of β from subjects' reaction times in the subsequent serial response task, as described above. We found that these two independent estimates of people's inverse temperatures are significantly correlated (excluding subjects for which $\beta = 0$ or $\beta \to \infty$), with a Spearman coefficient $r_s = 0.28$ ($p = 0.047$, permutation test). We note that we do not use the Pearson correlation coefficient because the estimates for β are not normally distributed for either the reaction time task ($p < 0.001$) nor the n-back task ($p = 0.013$) according to the Anderson-Darling test (*634*). This non-normality can be clearly seen in the distributions of β in Figs. 6.4a and 6.4g.

6.4.8 *Experimental procedures*

All participants provided informed consent in writing and experimental methods were approved by the Institutional Review Board of the University of Pennsylvania. In total, we recruited 634 unique participants to complete our studies on Amazon's Mechanical Turk. For the first serial response experiment, 101 participants only responded to sequences drawn from the modular graph, 113 participants only responded to sequences drawn from the lattice graph, and 72 participants responded to sequences drawn from both the modular and lattice graphs in back-to-back (counter-balanced) sessions for a total of 173 exposures to the modular graph and 185 exposures to the lattice graph.

For the second experiment, we recruited 120 subjects to respond to random walk sequences with Hamiltonian walks interspersed. For the third experiment, we recruited 78 participants to respond to sequences drawn from the ring graph with violations randomly interspersed. For the n-back experiment, 150 subjects performed the n-back task and, of those, 88 completed the subsequent serial response task. Worker IDs were used to exclude duplicate participants between experiments, and all participants were financially remunerated for their time. In the first experiment, subjects were paid up to $11 for up to an estimated 60 minutes: $3 per network for up to two networks, $2 per network for correctly responding on at least 90% of the trials, and $1 for completing the entire task. In the second and third experiments, subjects were paid up to $7.50 for an estimated 30 minutes: $5.50 for completing the experiment and $2 for correctly responding on at least 90% of the trials. In the n-back experiment, subjects were paid up to $8.50 for an estimated 45 minutes: $7 for completing the entire experiment and $1.50 for correctly responding on at least 90% of the serial response trials.

At the beginning of each experiment, subjects were provided with the following instructions: "In a few minutes, you will see five squares shown on the screen, which will light up as the experiment progresses. These squares correspond with keys on your keyboard, and your job is to watch the squares and press the corresponding key when that square lights up." For the 72 subjects that responded to both the modular and lattice graphs in the first experiment, an additional piece of information was also provided: "This part will take around 30 minutes, followed by a similar task which will take another 30 minutes." Before each experiment began, subjects were given a short quiz to verify that they had read and understood the instructions. If any questions were answered incorrectly, subjects were shown the instructions again and asked to repeat the quiz until they answered all questions correctly. Next, all subjects were shown a 10-trial segment that did not count towards their performance; this segment also displayed text on the screen explicitly telling the subject which keys to press on their keyboard. Subjects then began their 1500-trial experiment. For the subjects that responded to both the modular and lattice graphs, a brief reminder was presented before the second graph, but no new instructions were given. After completing each experiment, subjects were presented with performance information and their bonus earned, as well as the option to provide feedback.

6.5 SUPPLEMENTARY MATERIAL

In this Supplementary material, we provide extended discussion and data to support the results presented in the main text. The content is organized to roughly mirror the organization of the paper. In Sec. 6.5.1, we present experimental evidence that human reaction times – in addition to depending on higher-order network features – also reflect differences in fine-scale structure at the level of individual nodes. Just as for the higher-order effects presented in the main text, we demonstrate that these fine-scale phenomena are accurately predicted by our maximum entropy model. In Sec. 6.5.2, we present the mixed effects models that were used to estimate the cross-cluster

surprisal and modular-lattice effects. In Sec. 6.5.4, we demonstrate that the cross-cluster surprisal and modular-lattice effects cannot simply be explained by recency by directly controlling for the recency of stimuli. In Sec. 6.5.3, we use Hamiltonian walks to experimentally control for recency. In Sec. 6.5.5, we show that the cross-cluster surprisal and modular-lattice effects persist even when considering all 1500 trials for each subject. In Sec. 7.8, we show that the probability of an error on the serial response tasks increases for between- versus within-cluster transitions in the modular graph, indicating that the free energy framework can be used to predict human behaviors beyond reaction times. In Sec. 6.5.7, we present the mixed effects models that were used to estimate the effects of violations in the ring graph. In Sec. 6.5.8, we show that the effects of network violations cannot be explained by recency alone. In Sec. 6.5.9, we discuss why the forgetting of past stimuli altogether cannot explain the higher-order network effects that we examine in the main text. In Sec. 6.5.10, to aid in the reconstruction of our gradient descent algorithm for estimating the inverse temperature β from subjects' reaction times, we derive an analytic form for the gradient of the RMS prediction error of our model with respect to β. In Sec. 6.5.11, we discuss the relationship between our model and the successor representation in reinforcement learning.

6.5.1 *The effects of node heterogeneity on human expectations*

In the main text, we demonstrated that human expectations depend critically on the higher-order network structure of transitions. In addition to these higher-order phenomena, it has long been known that human expectations also reflect differences in the fine-scale structure of transition networks (*217*, *351*). For instance, humans are surprised by rare transitions, represented in a transition network by edges with low probability weight (*576*). Here, we provide empirical evidence showing that people's expectations also depend on the local topologies of the nodes that bookend a transition, and that these fine-scale effects are consistently predicted by our maximum entropy model.

In order to clearly study the effects of higher-order network structure, in the main text we focused on networks with uniform edge weights and node degrees. Here, to study the effects of node heterogeneity, we instead consider a set of Erdös-Rényi random graphs with the same number of nodes ($N = 15$) and edges (30) as in our previous modular and lattice graphs. To ensure that the random walks are properly defined, we set the transition probability A_{ij} of each edge in the graph to $1/k_i$, where k_i is the degree of node i. Since the probabilities A_{ij} decrease as the degree k_i increases, one should suspect that high-degree (or hub) nodes yield decreased anticipations – and therefore increased reaction times – at the next step of a random sequence. Indeed, using Eq. (6.6)), we find that our model analytically predicts decreased expectations following a high-degree node (Fig. 6.7a). Furthermore, across 177 human subjects, we find a strong positive correlation between people's reaction times and the degree of the preceding node in the sequence (Fig. 6.7b).

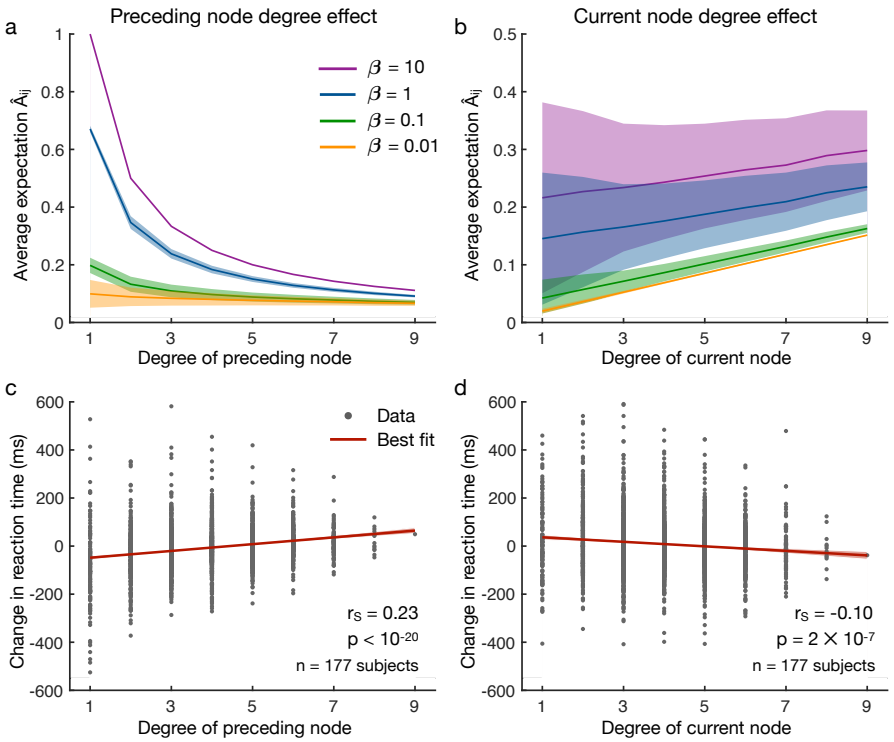


Figure 6.7: The effects of node degree on reaction times. (*a*) The average expectation $\hat{A}_{ij}$ plotted with respect to the degree of the preceding node i across a range of inverse temperatures β. As expected, expectations decrease as the degree of the preceding node increases; and for $\beta = 10$, we have $\hat{A}_{ij} \approx A_{ij} = 1/k_i$. The lines and shaded regions represent averages and 95% confidence intervals over 1000 randomly-generated Erdös-Rényi networks. (*b*) People exhibit sharp increases in reaction time following nodes of higher degree, with Spearman's correlation $r_S = 0.23$. The data is combined across 177 subjects, each of whom was asked to respond to a sequence of 1500 stimuli drawn from a random Erdös-Rényi network. Each data point represents the average reaction time for one node of a graph, and so each subject contributes 15 points. The line and shaded region represent the best fit and 95% confidence interval, respectively. (*c*) The average expectation $\hat{A}_{ij}$ plotted with respect to the degree of the current node j across the same range of inverse temperatures as in (*a*). (*d*) People exhibit a steady decline in reaction times as the current node degree increases, with Spearman's correlation $r_S = -0.10$. Source data are provided as a Source Data file.

Interestingly, while people's anticipations exhibit a sharp decline if the preceding node has high-degree, our model predicts that these hub nodes instead yield increased anticipations on the current step (Fig. 6.7c). Thus, while hub nodes give rise to marked increases in reaction times on the subsequent step, these high-degree nodes actually yield faster reactions on the current step (*351*) (Fig. 6.7d). This juxtaposition of effects from one time step to the next highlights the complex ways in which the network structure of transitions can affect people's mental representations. Additionally, the success of our model in predicting these competing phenomena further strengthens our conclusion that mental errors play a crucial role in shaping people's internal expectations.

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1418.9 ± 73.1	19.42	< 0.001	$***$
log(Trial)	-92.1 ± 9.2	-9.96	< 0.001	$***$
Stage	-551.5 ± 85.0	-6.48	< 0.001	$***$
Recency	1.4 ± 0.1	23.57	< 0.001	$***$
Trans_Type	34.9 ± 6.0	5.77	< 0.001	$***$
log(Trial):Stage	67.0 ± 11.4	5.89	< 0.001	$***$

Table 6.1: Mixed effects model measuring the cross-cluster surprisal effect. A mixed effects model fit to the reaction time data for the modular graph with the goal of measuring the cross-cluster surprisal effect. We find a significant 35 ms increase in reaction times (173 subjects) for between-cluster transitions versus within-cluster transitions (grey). The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

6.5.2 *Measuring higher-order network effects*

In order to extract the effects of higher-order network structure on subjects' reaction times, we use linear mixed effects models, which have become prominent in human research where many measurements are made for each subject (*39*, *582*). To fit our mixed effects models and to estimate the statistical significance of each effect we use the `fitlme` function in MATLAB (R2018a). In what follows, when referring to our mixed effects models, we adopt the standard R notation (*65*).

6.5.2.1 *Cross-cluster surprisal effect*

We first measure the cross-cluster surprisal effect (Fig. 6.2a) using a mixed effects model with the formula '$\text{RT} \sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Trans\_Type} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} + \text{Trans\_Type}\,|\,\text{ID})$', where RT is the reaction time, Trial is the trial number between 501 and 1500, Stage is the stage of the experiment (either one or two), Target is the target button combination, Recency is the number of trials since last observing a node (*41*), Trans_Type is the type of transition (either within-cluster or between-cluster), and ID is each subject's unique ID. We remark that our inclusion of Recency in the model is intended to distinguish the graph effects that we are interested in studying from the possible confound of recency, an effect that we directly control for in Sec. 6.5.4. The mixed effects model is summarized in Tab. 7.1, reporting a 35 ms increase in reaction times for between-cluster transitions relative to within-cluster transitions (Fig. 6.2a). This result is measured from the reaction time data for all 173 subjects that observed random walks in the modular graph.

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1436.5 ± 48.4	29.67	< 0.001	$***$
log(Trial)	-97.2 ± 6.1	-15.89	< 0.001	$***$
Stage	-555.2 ± 59.3	-9.36	< 0.001	$***$
Recency	1.7 ± 0.1	29.94	< 0.001	$***$
Graph	22.8 ± 5.8	3.95	< 0.001	$***$
log(Trial):Stage	71.4 ± 8.4	8.48	< 0.001	$***$

Table 6.2: Mixed effects model measuring the modular-lattice effect. A mixed effects model fit to the reaction time data for the modular and lattice graphs with the goal of measuring the modular-lattice effect. We find a significant 23 ms increase in reaction times overall (72 subjects) in the lattice graph relative to the modular graph (grey). The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1420.8 ± 162.3	8.75	< 0.001	$***$
log(Trial)	-101.4 ± 22.7	-4.48	< 0.001	$***$
Recency	0.6 ± 0.1	5.00	< 0.001	$***$
Trans_Type	35.6 ± 13.7	2.59	0.010	$**$

Table 6.3: Mixed effects model measuring the cross-cluster surprisal effect in Hamiltonian walks. A mixed effects model fit to subjects' reaction times in Hamiltonian walks on the modular graph with the goal of measuring the cross-cluster surprisal effect. We find a significant 36 ms increase in reaction times (120 subjects) for between-cluster transitions versus within-cluster transitions (grey). The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

6.5.2.2 *Modular-lattice effect*

We next measure the modular-lattice effect (Fig. 6.2b) using a mixed effects model of the form 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Graph} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \,|\, \text{ID})$', where Graph represents the type of transition network, either modular or lattice. Note that we only include Graph as a fixed effect because the corresponding mixed effect is not statistically significant. The mixed effects model is summarized in Tab. 7.2, reporting a 23 ms increase in reaction times in the lattice graph relative to the modular graph (Fig. 6.2b). This result is measured from the reaction time data for the 72 subjects that observed random walks in both the modular and lattice graphs.

6.5.3 *Cross-cluster surprisal with Hamiltonian walks*

Throughout the main text, we assume that people's reaction times reflect their internal representations of the transition structure. To justify this assumption, we must show

that the higher-order network effects cannot simply be explained by recency. Here, we measure the cross-cluster surprisal effect while experimentally controlling for recency using Hamiltonian walks. In contrast to random walks, Hamiltonian walks visit each node in the transition graph exactly once, thereby guaranteeing that each node is visited once every 15 trials. We run a new experiment in which each subject (out of 120 subjects) is presented with a sequence of 1500 stimuli drawn from the modular graph: The first 700 nodes reflect a standard random walk, while the remaining 800 trials consist of 8 repeated segments of 85 stimuli specified by a random walk followed by 15 stimuli specified by a Hamiltonian walk. The initial 700 random walk trials are meant to constitute a learning phase in which the subject builds an internal representation of the modular graph. Since, in the modular graph, Hamiltonian walks do not obey the same transition probabilities as random walks, the sequences of 85 random walk trials between each Hamiltonian sequence are meant to help the subject maintain their learned representation. Within the set of Hamiltonian walks through the modular graph, the probability of transitioning from one cluster boundary node to the adjacent one (if not already visited) is 1, whereas the probability of transitioning from the latter boundary node to each of the adjacent non-boundary nodes is $1/3$. To eliminate this difference, we randomly selected one fixed Hamiltonian walk for each subject. This fixed walk was entered at a different node depending on where the preceding walk terminated, and we randomly switched between forward and backward traversals for each walk (*584*).

We measure the cross-cluster surprisal within the Hamiltonian trials using a mixed effects model with the formula 'RT $\sim$ log(Trial) + Target + Recency + Trans_Type + (1 + log(Trial) + Recency + Trans_Type|ID)', where each of the variables has been defined previously. The model is summarized in Tab. 7.3, reporting a 36 ms increase in reaction times for between-cluster transitions relative to within-cluster transitions within Hamiltonian trials ($p = 0.010$), matching (within errors) the effect size reported in the original experiment that only included random walks (see Tab. 7.1). This result is measured from the reaction time data for all 120 subjects that observed random walks with Hamiltonian walks interspersed in the modular graph. This result indicates that the cross-cluster surprisal effect cannot be explained by recency alone, and must therefore must be at least partially driven by people's internal representations of the transition structure.

6.5.3.1 *Removing Hamiltonian trials before the first cross-cluster transition*

The purpose of the Hamiltonian walk experiment described above is to experimentally control for the effects of recency on people's reaction times. However, when thinking carefully about the transition from a random walk to a Hamiltonian walk, it becomes clear that recency might still have a noticeable impact. Consider, for example the last few trials of a random walk preceding a transition to a Hamiltonian walk – the corresponding stimuli are likely to belong to the same module in the modular graph. When the sequence converts to a Hamiltonian walk, the first few stimuli are also

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1536.6 ± 178.0	8.63	< 0.001	$***$
log(Trial)	-116.3 ± 24.8	-4.68	< 0.001	$***$
Recency	0.4 ± 0.1	3.00	0.003	$**$
Trans_Type	28.2 ± 13.9	2.03	0.043	$*$

Table 6.4: Mixed effects model measuring the cross-cluster surprisal effect in restricted Hamiltonian walks. A mixed effects model fit to subjects' reaction times after the first cross-cluster transition within each Hamiltonian walk. We find a significant 28 ms increase in reaction times (120 subjects) for between-cluster transitions versus within-cluster transitions (grey). The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

likely to belong to the same module, thereby inducing a decrease in reaction times due to recency. Therefore, in order to more thoroughly control for recency effects, we considered only trials after the first cross-cluster transition within each Hamiltonian walk. We carry out this restricted analysis using the same form for the mixed effects model as that described above: 'RT $\sim$ log(Trial) + Target + Recency + Trans_Type + (1 + log(Trial) + Recency + Trans_Type | ID)'. The model, which is summarized in Tab. 7.4, estimates a significant cross-cluster surprisal effect of 28 ms ($p = 0.043$), again matching within errors the effect size found in the original random walk data.

6.5.3.2 *Decreasing cross-cluster surprisal with increasing Hamiltonian trials*

As discussed above, the first 700 trials of each sequence were drawn from a random walk to allow subjects to build an internal representation of the random walk transition structure. Since the transition probabilities reflected in the Hamiltonian walks differ from those in the random walks, we expect subjects' representations of the transition structure to shift as they observe increasing numbers of Hamiltonian trials. Therefore, to further establish the notion that people's reactions are primarily driven by their internal representations, here we show that the strength of the cross-cluster surprisal decreases as subjects observe increasing numbers of Hamiltonian trials. To do so, we use a mixed effects model with the formula 'RT $\sim$ log(Trial) $*$ Trans_Type + Target + Recency + (1 + log(Trial) + Recency + Trans_Type | ID)', where the only difference with the formula above is that here we include an interaction term between log(Trial) and Trans_Type. The results of the fitted model are summarized in Tab. 7.5, reporting a significant decrease in the strength of the cross-cluster surprisal with increasing Hamiltonian trials ($p = 0.024$).

6.5.3.3 *Experimental setup and procedures*

Subjects performed a self-paced serial reaction time task, as described in the Methods section of the main text. The only difference between this experiment and the original random walk experiments is that the 1500 trials were split into 700 trials drawn as a

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1394.1 ± 188.8	7.39	< 0.001	$***$
log(Trial)	-96.1 ± 26.4	-3.64	< 0.001	$***$
Recency	0.4 ± 0.1	3.02	0.003	$**$
Trans_Type	640.3 ± 271.9	2.35	0.019	$*$
log(Trial):Trans_Type	-87.2 ± 38.7	-2.25	0.024	$*$

Table 6.5: Mixed effects model measuring the decrease in cross-cluster surprisal with increasing Hamiltonian trials. A mixed effects model fit to subjects' reaction times in Hamiltonian walks on the modular graph with the goal of measuring the dependence of the cross-cluster surprisal on increasing trial number. We find a significant decrease in the strength of the cross-cluster surprisal with increasing trials (grey), indicating that the introduction of Hamiltonian walks weakens people's internal representations of the random walk structure (120 subjects). The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

random walk and a subsequent 800 trials divided into 8 segments of 85 random walk trials followed by 15 Hamiltonian walk trials, all drawn from the modular graph. In total, we recruited 120 subjects to perform this Hamiltonian walk experiment, and they were paid up to $5 each for an estimated 30 minutes: $3.50 for completing the task and $1.50 for correctly responding on at least 90% of the trials.

6.5.4 *Controlling for recency in random walks*

In the previous section, we showed that cross-cluster surprisal remains significant during Hamiltonian walks, which experimentally control for the recency of stimuli. Building upon this result, in this section we measure the cross-cluster surprisal and modular-lattice effects in our initial random walk data while filtering our data based on stimulus recency.

6.5.4.1 *Cross-cluster surprisal effect while controlling for recency*

In order to control for recency, we filter our data to only include trials in which the current stimulus was last seen a specific number of trials in the past. For example, when studying trials with a recency of four, we only consider reaction times from our experiments for which the current stimulus was last seen four trials previously. After filtering the data, we then estimate the cross-cluster surprisal effect using a mixed effects model of the form 'RT $\sim$ log(Trial) $*$ Stage + Target + Trans_Type + (1 + log(Trial) $*$ Stage + Trans_Type | ID)'. Fig. 6.8a shows the estimated increase in reaction times for within-cluster versus between-cluster transitions after controlling for recency. Specifically, we consider recency values of two (the minimum) through nine, and we also consider trials with recency greater than or equal to 10, for which the effects of recency should be small. We remark that we do not include trials of recency three in

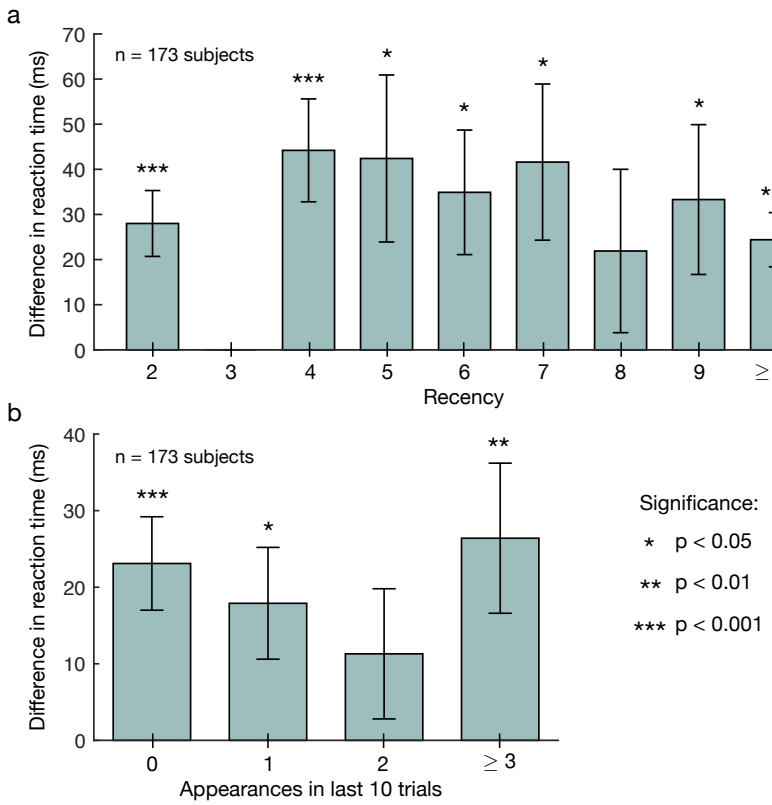


Figure 6.8: Cross-cluster surprisal while controlling for recency. (*a*) Increase in reaction times for between-cluster versus within-cluster transitions in the modular graph after controlling for the recency of stimuli. We note that, due to the topology of the modular graph, there do not exist between-cluster transitions with recency three. We find significant cross-cluster surprisal effects for all recency values besides eight. (*b*) Increase in reaction times for between- versus within-cluster transitions after controlling for the number of times that the current stimulus has appeared in the previous 10 trials. We observe significant cross-cluster surprisal for all numbers of recent stimulus appearances besides two. Effect sizes (represented by bar plots), standard deviations (represented by error bars), and *p*-values are estimated using mixed effects models. The results are measured for all 173 subjects that observed random walks in the modular graph. Source data are provided as a Source Data file.

our analysis because, due to the topology of the modular graph, there do not exist between-cluster transitions with recency three. We find significant effects for all recency values besides eight.

In addition to controlling for the recency of stimuli, we also study the cross-cluster surprisal while controlling for the number of appearances of the current stimulus in the last 10 trials. In particular, we filter our data to only include trials for which the current stimulus was seen a specified number of times in the previous 10 trials, and for each set of filtered data we estimate the cross-cluster surprisal using a mixed effects model of the same form as above. We observe a significant increase in reaction times for between- versus within-cluster transitions for all trials except for those for which the stimulus appeared twice in the last 10 trials (Fig. 6.8b). Together, these results

demonstrate that the cross-cluster surprisal effect cannot be explained by recency alone, and therefore must stem, at least in part, from people's internal representations of the transition structure.

6.5.4.2 *Modular-lattice effect while controlling for recency*

We next consider the modular-lattice effect after controlling for recency. Filtering the data from the modular and lattice graphs to only include trials of a given recency, we estimate the difference in reaction times between the two graphs using a mixed effects model of the form 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Graph} + (1 + \log(\text{Trial}) * \text{Stage}|\text{ID})$'. Fig. 6.9a shows that we find a significant increase in reaction times for the lattice graph relative to the modular graph for all recency values considered besides three, nine, and $\geqslant 10$. Additionally, in Fig. 6.9b, we control for the number of appearances of the current stimulus in the previous 10 trials. Using a mixed effects model of the same form as that above, we find a significant modular-lattice effect in two of the four conditions. Together, these results demonstrate that the difference in reaction times between the modular and lattice graphs persists after controlling for the recency of stimuli, indicating that people are better able to anticipate transitions in the modular graph than in the lattice graph.

6.5.5 *Measuring network effects including early trials*

Throughout the above analysis of the serial response tasks, we purposefully omitted the first 500 trials for each subject, choosing instead to focus on the final 1000 trials. We did this in order to allow the subjects to build an internal representation of each network structure before probing their anticipations of transitions. Here, we show that this data processing step is not necessary to observe higher-order network effects; that is, we show that there exist significant network effects even if we include the first 500 trials in our analysis.

6.5.5.1 *Cross-cluster surprisal effect with early trials*

We first consider the cross-cluster surprisal effect defined by an increase in reaction times for transitions between clusters relative to transitions within clusters in the modular graph. Using a mixed effects model of the same form as that used in the previous analysis in Sec. 6.5.2 (i.e., 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Trans\_Type} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} + \text{Trans\_Type}|\text{ID})$'), and including all 1500 trials for each subject, we find a significant 35 ms increase in reaction times for between- versus within-cluster transitions (Tab. 7.6). We note that this effect is even larger than that observed in our previous analysis (Tab. 7.1).

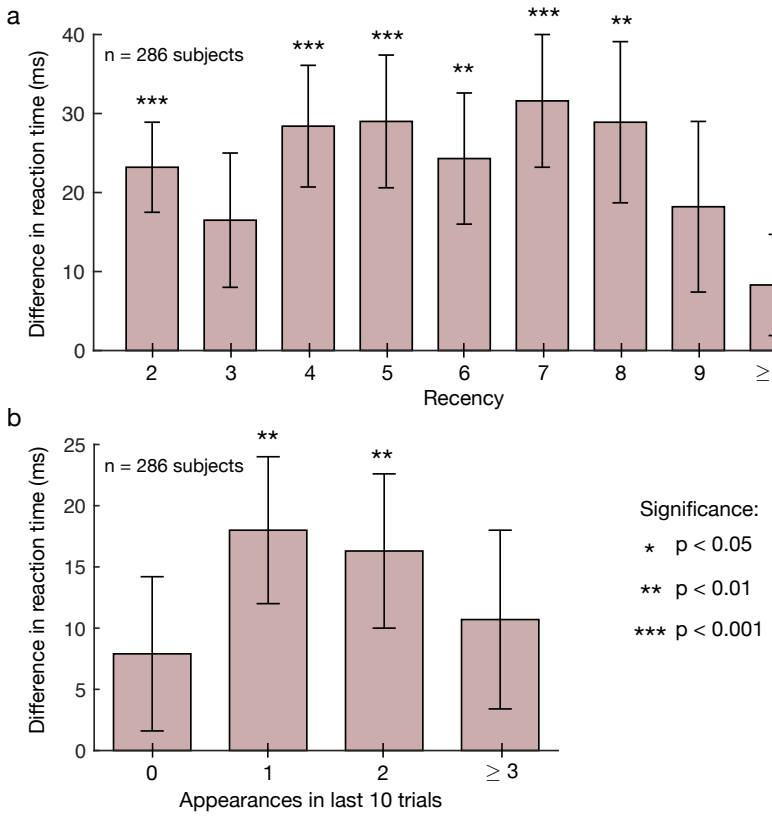


Figure 6.9: Modular-lattice effect while controlling for recency. (*a*) Difference in reaction times between the lattice and modular graphs after controlling for the recency of stimuli. We observe a significant increase in reaction times for the lattice graph relative to the modular graph for all recency values besides three, nine, and $\geqslant$ 10. (*b*) Difference in reaction times between the lattice and modular graphs after controlling for the number of times the current stimulus has appeared in the previous 10 trials. We find a significant modular-lattice effect for one and two stimulus appearances in the last 10 trials. Effect sizes (represented by bar plots), standard deviations (represented by error bars), and *p*-values are estimated using mixed effects models. The results are measured for all 72 subjects that observed random walks in both the modular and lattice graphs. Source data are provided as a Source Data file.

6.5.5.2 *Modular-lattice effect with early trials*

We next consider the modular-lattice effect defined by an increase in reaction times in the lattice graph relative to the modular graph. Using a mixed effects model of the same form as that used in the previous analysis in Sec. 6.5.2 (i.e., 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Graph} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \,|\, \text{ID})$'), and including all 1500 trials for each subject, we find a significant 16 ms increase in reaction times in the lattice versus the modular graph (Tab. 7.7). These results demonstrate that higher-order network effects studied in the main text exist throughout the entire serial response task.

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1340.0 ± 44.4	30.19	< 0.001	$***$
log(Trial)	-88.7 ± 5.0	-17.04	< 0.001	$***$
Stage	-473.1 ± 47.6	-9.93	< 0.001	$***$
Recency	1.5 ± 0.1	24.65	< 0.001	$***$
Trans_Type	35.4 ± 6.0	5.94	< 0.001	$***$
log(Trial):Stage	60.4 ± 5.5	11.06	< 0.001	$***$

Table 6.6: Mixed effects model measuring the cross-cluster surprisal effect including the first 500 trials. A mixed effects model fit to all of the reaction time data, including the first 500 trials for each subject, for the modular graph with the goal of measuring the cross-cluster surprisal effect. We find a significant 35 ms increase in reaction times (173 subjects) for between-cluster transitions versus within-cluster transitions. The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1357.0 ± 30.3	44.79	< 0.001	$***$
log(Trial)	-87.8 ± 3.4	-26.06	< 0.001	$***$
Stage	-490.7 ± 25.3	-19.38	< 0.001	$***$
Recency	2.0 ± 0.1	32.35	< 0.001	$***$
Graph	16.3 ± 5.4	3.00	0.003	$**$
log(Trial):Stage	62.7 ± 3.5	17.76	< 0.001	$***$

Table 6.7: Mixed effects model measuring the modular-lattice effect including the first 500 trials. A mixed effects model fit to all of the reaction time data, including the first 500 trials for each subject, for the modular and lattice graphs with the goal of measuring the modular-lattice effect. We find a significant 16 ms increase in reaction times overall (72 subjects) in the lattice graph relative to the modular graph. The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

6.5.6 *Network effects on error trials*

Thus far we have focused on predicting human reaction times as a proxy for people's anticipations of transitions. Another way to probe anticipation is by studying the trials on which subjects respond incorrectly; one might expect that the probability of an erroneous response should increase with decreasing anticipation. Here, we test this hypothesis for between- versus within-cluster transitions in the modular graph and for all transitions in the modular graph versus the lattice graph.

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.005 ± 0.012	0.39	0.697	
log(Trial)	0.004 ± 0.002	2.14	0.032	$*$
Stage	0.015 ± 0.007	2.14	0.032	$*$
Recency	< 0.001	16.52	< 0.001	$***$
Trans_Type	0.004 ± 0.002	2.83	0.005	$**$

Table 6.8: Mixed effects model measuring the cross-cluster effect on task errors. A mixed effects model fit to predict error trials for the modular graph with the goal of measuring the cross-cluster effect on task errors. We find a significant increase in task errors (173 subjects) for between-cluster transitions relative to within-cluster transitions (grey). The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

6.5.6.1 *Cross-cluster surprisal effect on errors*

First, we consider the cross-cluster surprisal effect on errors defined by an increase in task errors for transitions between clusters relative to transitions within clusters in the modular graph. We employ a mixed effects model with formula 'Error $\sim$ log(Trial) + Stage + Target + Recency + Trans_Type + (1 + log(Trial) | ID)', where Error indicates whether the subject provided an incorrect ('1') or correct ('0') response. Note that, relative to our measurement of the cross-cluster surprisal for reaction times in Sec. 6.5.2, we have removed the fixed effect interaction between log(Trial) and Stage as well as the mixed effects for the variables Stage, Recency, and Trans_Type because they are not statistically significant in this setting. We find a significant increase in errors for between- versus within-cluster transitions (Tab. 7.8), suggesting yet again that subjects have weaker anticipation for cross-cluster transitions than for within-cluster transitions.

6.5.6.2 *Modular-lattice effect on errors*

Second, we consider the modular-lattice effect on errors defined by an increase in task errors for the lattice graph relative to the modular graph. We employ a mixed effects model with formula 'Error $\sim$ log(Trial) + Stage + Target + Recency + Graph + (1 + log(Trial) + Recency + Graph | ID)', where each of the variables has been defined previously. We again note that we have removed the interaction between log(Trial) and Stage because it was not statistically significant in our prediction of task errors. Inspecting the mixed effects model described in Tab. 7.9, we do not find a significant difference in the number of task errors between the modular and lattice graphs. One possible explanation for this lack of an effect is that people's task accuracy is predominantly impacted by very poorly anticipated transitions. Thus, while anticipation in the lattice graph is lower than that in the modular graph on average, it could be the case that the significant decrease in anticipation for cross-cluster transitions in the modular graph yields a similar number of task errors overall.

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.026 ± 0.009	3.05	0.002	**
log(Trial)	0.002 ± 0.001	1.47	0.142	
Stage	0.003 ± 0.003	0.98	0.325	
Recency	< 0.001	14.62	< 0.001	***
Graph	-0.004 ± 0.003	-1.34	0.180	

Table 6.9: Mixed effects model measuring the modular-lattice effect on task errors. A mixed effects model fit to predict error trials for the modular and lattice graphs with the goal of measuring the modular-lattice effect on task errors. We do not find a significant change in errors based on the graph (grey; 72 subjects). The significance column represents p-values less than 0.001 (***), less than 0.01 (**), and less than 0.05 (*). Source data are provided as a Source Data file.

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1352.7 ± 79.2	17.07	< 0.001	***
log(Trial)	-101.1 ± 10.2	-9.96	< 0.001	***
Recency	2.1 ± 0.1	16.20	< 0.001	***
Top_Dist (short vs. no violation)	37.9 ± 8.4	4.50	< 0.001	***
Top_Dist (long vs. no violation)	63.3 ± 7.8	8.07	< 0.001	***

Table 6.10: Mixed effects model measuring the effects of violations relative to standard transitions. A mixed effects model fit to the reaction time data for the ring graph with the goal of measuring the effects of violations relative to standard transitions. We find a significant increase in reaction times of 38 ms (78 subjects) for short violations and 63 ms for long violations (grey), even after accounting for recency effects. The significance column represents p-values less than 0.001 (***), less than 0.01 (**), and less than 0.05 (*). Source data are provided as a Source Data file.

6.5.7 *Measuring the effects of network violations*

We study the effects of violations of varying topological distance in the ring graph using a mixed effects model with the formula 'RT $\sim$ log(Trial) + Target + Recency + Top_Dist + (1 + log(Trial) + Recency + Top_Dist | ID)', where Top_Dist represents the topological distance of a transition, either one for a standard transition, two for a short violation, or three for a long violation. The results of fitting this mixed effects model are summarized in Tab. 7.10, reporting increases in reaction times over standard transitions of 38 ms for short violations and 63 ms for long violations. Second, to measure the difference in reaction times between long and short violations, we implemented a model of the same form, but restricted Top_Dist to only include short violations of topological distance two and long violations of topological distances three and four. This model is summarized in Tab. 7.11, reporting a 28 ms increase in reaction times for

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1380.9 ± 156.1	8.84	< 0.001	$***$
log(Trial)	-97.1 ± 21.3	-4.57	< 0.001	$***$
Recency	0.7 ± 0.3	2.67	0.008	$**$
Top_Dist (long vs. short violation)	28.4 ± 11.2	2.54	0.011	$*$

Table 6.11: Mixed effects model measuring the effects of long versus short violations. A mixed effects model fit to the reaction time data for the ring graph with the goal of measuring the effects of long versus short violations. We find a significant 28 ms increase in reaction times (78 subjects) for long violations relative to short violations (grey), even after accounting for recency effects. The significance column represents p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$). Source data are provided as a Source Data file.

long violations relative to short violations. This result is measured from all 78 subjects that observed random walks with violations in the ring graph.

6.5.8 *Controlling for recency: Network violations*

In the main text, we attribute the observed increase in reaction times for network violations to subjects' internal representations of the transition structure. Alternatively, these effects could be due to the fact that standard transitions are more likely than network violations to yield a stimulus that has been seen recently. To show that the effects of network violations are not simply driven by recency, we directly control for the recency of stimuli in our data. Because the violations data is more sparse than the standard random walk data (we only observe 50 violations per subject, split between 20 short violations and 30 long violations), and because the network violations often yield stimuli with large recency values (for example, 69% of violations yield stimuli with recency greater than 10), we separate our data based on ranges of recency values that provide an approximately even distribution of violations (see Fig. 6.10a). After separating the data by recency, we estimate the effects of short and long violations using a mixed effects model of the form 'RT $\sim$ log(Trial) + Target + Top_Dist + (1 + log(Trial) | ID)'. We note that, in comparison to the model used in Sec. 6.5.7, we have removed the mixed effect of Top_Dist because the filtered datasets are not large enough to provide a significant estimate. In Fig. 6.10a, we see that, within each recency range other than recency greater than 40, at least one of either the short or long violations generates a significant increase in reaction times relative to standard transitions. Additionally, in Fig. 6.10b, we filter the violations data by the number of appearances of the current stimulus in the previous 10 trials. Network violations yield significant increases in reaction times across all conditions other than short violations with zero appearances in the last 10 trials. Together, these results demonstrate that the effects of network violations cannot simply be explained by recency, therefore suggesting that subjects maintain an internal representation of the transition structure.

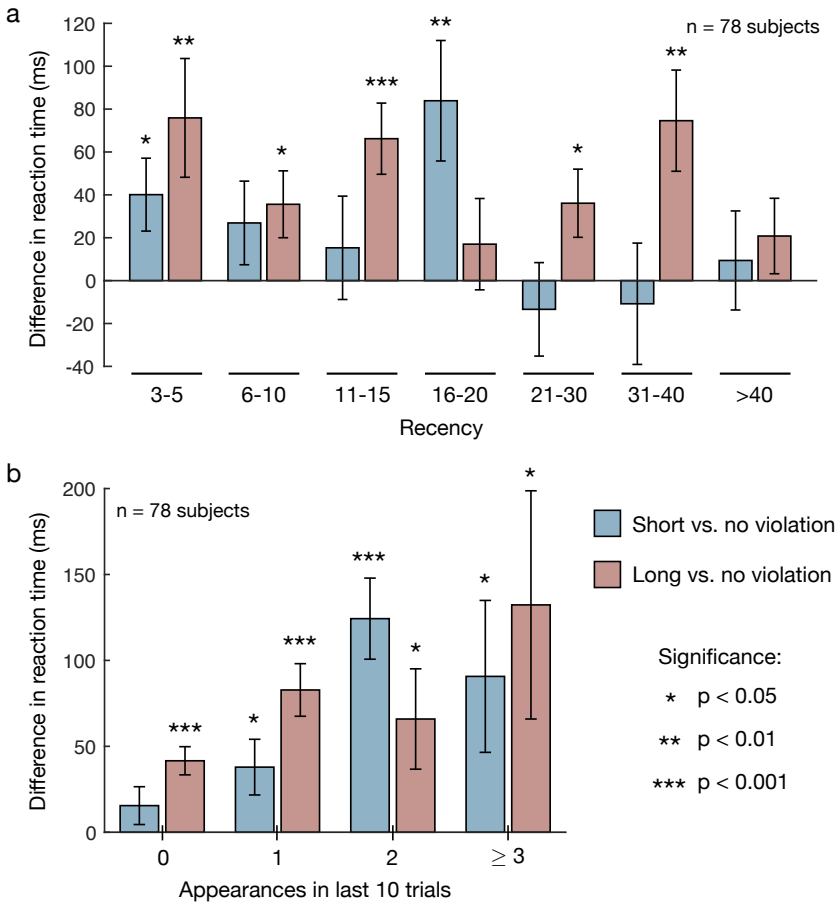


Figure 6.10: Comparing standard transitions to network violations while controlling for recency. (*a*) Difference in reaction times between standard transitions and short violations (blue) or long violations (red) in the ring graph after controlling for the recency of stimuli. We observe at least one significant effect of network violations for all recency ranges less than 40. (*b*) Increase in reaction times for short (blue) and long (red) network violations after controlling for the number of times the current stimulus has appeared in the previous 10 trials. For long violations, we find a significant increase in reaction times across all numbers of recent stimulus appearances. For short violations, we find a significant increase in reaction times across all numbers of recent stimulus appearances besides zero. Effect sizes (represented by bar plots), standard deviations (represented by error bars), and p-values are estimated using mixed effects models. The results are measured for all 78 subjects that observed random walks with violations in the ring graph. Source data are provided as a Source Data file.

We repeat the above analysis to measure the difference in reaction times between short and long violations while controlling for recency. We observe a significant increase in reaction times for long violations relative to short violations in two of the seven recency ranges (Fig. 6.11a). However, we do not report a significant difference in reaction times after controlling for the number of appearances of stimuli in the previous 10 trials (Fig. 6.11b). We remark that, given the noisy nature of reaction times and the small number of measurements per subject, the large standard deviations in

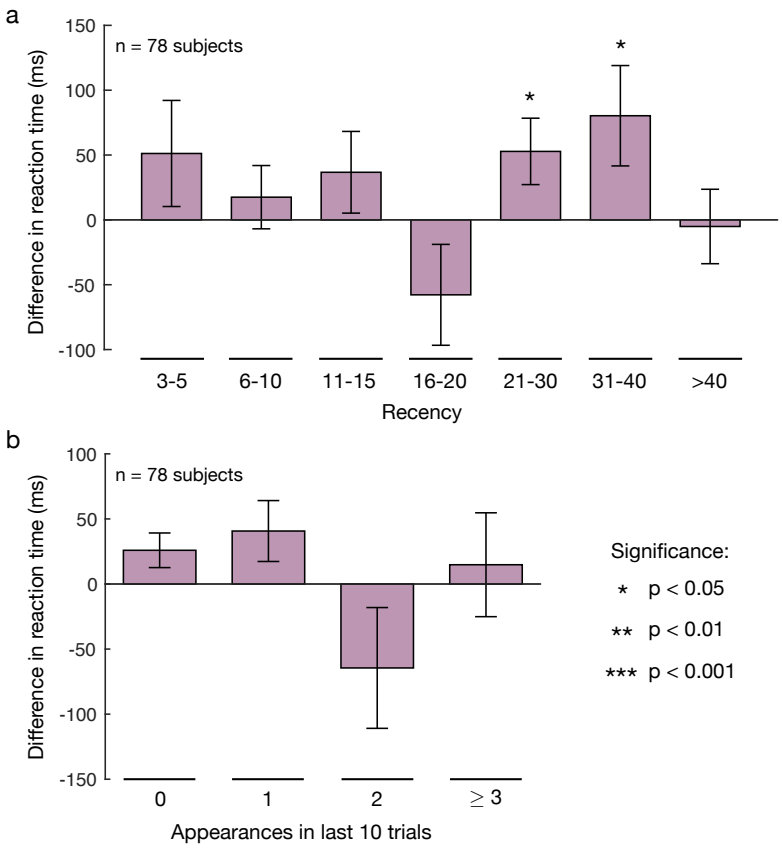


Figure 6.11: Comparing short versus long network violations while controlling for recency. (*a*) Difference in reaction times between short and long network violations after controlling for the recency of stimuli. We find significant increases in reaction times for long violations in the recency ranges 21-30 and 31-40. (*b*) Difference in reaction times between short and long network violations after controlling for the number of times the current stimulus has appeared in the previous 10 trials. Effect sizes (represented by bar plots), standard deviations (represented by error bars), and *p*-values are estimated using mixed effects models. The results are measured for all 78 subjects that observed random walks with violations in the ring graph. Source data are provided as a Source Data file.

our estimates are not surprising. Nevertheless, these results provide tentative evidence that recency alone cannot explain the difference in reaction times between long and short network violations.

6.5.9 *The forgetting of stimuli cannot explain network effects*

In the derivation of our model, the central mathematical object is the memory distribution $P(\Delta t)$, which represents the probability that a person recalls the stimulus that occurred at time $t - \Delta t$ instead of the stimulus that they were trying to recall at time t. Generally, this memory distribution is intended to reflect the erroneous shuffling of past stimuli in a person's memory. Alternatively, one could imagine errors in memory

that reflect the forgetting of past stimuli altogether, a process that has recently been shown to impact human reinforcement learning (*153*, *154*) and to facilitate flexible and generalizable decision making (*555*). Here we argue that this second form of cognitive errors – that is, the simple forgetting of stimuli – cannot explain the higher-order network effects described in the main text.

Consider a sequence of stimuli reflecting a random walk of length T on a network defined by the transition matrix A, where A_{ij} represents the probability of transitioning from stimulus i to stimulus j. Given a running tally $n_{ij}(T)$ of the number of times each transition has occurred, we recall that the most accurate prediction for the transition structure is given by the maximum likelihood estimate $\hat{A}^{\text{MLE}}_{ij}(T) = n_{ij}(T) / \sum_k n_{ik}(T)$. Now suppose that a human learner forgets each observed transition at some fixed rate. On average, this process of estimating A after forgetting some number of transitions uniformly at random is equivalent to estimating A at some prior time T_{eff}. In other words, forgetting observed transitions at random simply introduces additional white noise into the transitions estimates $\hat{A}^{\text{MLE}}_{ij}(T)$. As discussed in the main text, maximum likelihood estimation provides an unbiased estimate of the transition structure, and therefore cannot explain the fact that people's representations depend systematically on higher-order network organization. Similarly, the addition of white noise to $\hat{A}^{\text{MLE}}(T)$ will also yield an unbiased (although less accurate) estimate of the transition structure. Therefore, while the forgetting of past stimuli plays an important role in a number of cognitive processes (*153*, *154*, *555*), this mechanism cannot be used to explain the higher-order network effects observed in human experiments and predicted by our model.

6.5.10 *Gradient of RMS error with respect to inverse temperature* β

Given a sequence of nodes x_t, we recall that our prediction for the reaction time at time t is given by $\hat{r}(t) = r_0 + r_1 a(t)$, where $a(t) = \hat{A}_{x_{t-1}, x_t}(t-1)$ is the predicted anticipation of node x_t. The gradient of the RMS error $\text{RMSE} = \sqrt{\frac{1}{T}\sum_t (r(t) - \hat{r}(t))^2}$ with respect to the inverse temperature β is given by

$$\frac{\partial \text{RMSE}}{\partial \beta} = \frac{-r_1}{T} \frac{1}{\text{RMSE}} \sum_t \left(r(t) - \hat{r}(t)\right) \frac{\partial a(t)}{\partial \beta}, \tag{6.12}$$

where the derivative of the anticipation is given by

$$\frac{\partial \hat{A}_{ij}(t)}{\partial \beta} = \frac{1}{\sum_k \tilde{n}_{ik}(t)} \left(\frac{\partial \tilde{n}_{ij}(t)}{\partial \beta} - \hat{A}_{ij}(t) \sum_\ell \frac{\partial \tilde{n}_{i\ell}(t)}{\partial \beta} \right). \tag{6.13}$$

Recalling Eq. (6.8), the derivative of the transition counts can be written

$$\frac{\partial \tilde{n}_{ij}(t)}{\partial \beta} = \sum_{t'=1}^{t-1} \sum_{\Delta t=0}^{t'-1} \frac{\partial P_{t'}(\Delta t)}{\partial \beta} \left[i = x_{t'-\Delta t}\right] \left[j = x_{t'+1}\right], \tag{6.14}$$

where $P_{t'}(\Delta t)$ represents the probability of accidentally remembering the node $x_{t'-\Delta t}$ instead of the target node $x_{t'}$. Taking one more derivative, we have

$$\frac{\partial P_{t'}(\Delta t)}{\partial \beta} = P_{t'}(\Delta t)\left(-\Delta t + \sum_{\Delta t'=0}^{t'-1} P_{t'}(\Delta t')\Delta t' \right). \tag{6.15}$$

Taken together, Eqs. (6.12-6.15) define the derivative of the RMS error with respect to the inverse temperature β, thus completing the description of our gradient descent algorithm.

6.5.11 *Connection to the successor representation*

In the limit of an infinitely-long sequence of nodes, we showed in the main text that the transition estimates in our model take the following concise analytic form,

$$\hat{A} = (1 - e^{-\beta})A(I - e^{-\beta}A)^{-1}, \tag{6.16}$$

where A is the true transition structure, β is the inverse temperature in our memory distribution, and I is the identity matrix. Interestingly, this equation takes a similar form to the successor representation from reinforcement learning,

$$M = A(I - \gamma A)^{-1}, \tag{6.17}$$

where γ is the future discount factor, which tunes the desired time-scale over which a person wishes to make predictions (*247*, *645*). Put simply, starting at some node i, the successor representation M_{ij} counts the future discounted occupancy of node j. Identifying $\gamma = e^{-\beta}$, we notice that the successor representation is equivalent to an unnormalized version of our transition estimates. Moreover, the same mathematical form crops up in complex network theory, where it is known as the communicability between nodes in a graph (*209*, *210*, *242*).

The relationship between the transition estimates in our model and the successor representation is fascinating, especially given the marked differences in the concepts that the two models are based upon. In our model, people attempting to learn the one-step transition structure A instead arrive at the erroneous estimate $\hat{A}$ due to natural errors in perception and recall. By contrast, given a desired time-scale γ, the successor representation defines the optimal prediction of node occupancies into the future (*247*, *645*). Interestingly, the successor representation has been linked to grid cells and abstract representations in the hippocampus (*242*, *629*), decision making in reward-based tasks (*451*, *573*), and the temporal difference and temporal context models of learning and memory (*247*, *324*, *645*). The successor representation assumes that humans are making predictions multiple steps into the future; however, our results show that a similar mathematical form can instead represent a person who simply attempts to predict one step into the future, but misses the mark due to natural errors

in cognition. This biologically-plausible hypothesis of erroneous predictions highlights the importance of thinking carefully about the impact of mental errors on human learning (*153*, *154*, *555*).

7

HUMAN INFORMATION PROCESSING IN COMPLEX NETWORKS

This chapter contains work from Lynn, Christopher W., Lia Papadopoulos, Ari E. Kahn, and Danielle S. Bassett. "Human information processing in complex networks." Nature Physics, in press.

Abstract

Humans communicate using systems of interconnected stimuli or concepts – from language and music to literature and science – yet it remains unclear how, if at all, the structure of these networks supports the communication of information. Although information theory provides tools to quantify the information produced by a system, traditional metrics do not account for the inefficient ways that humans process this information. Here we develop an analytical framework to study the information generated by a system as perceived by a human observer. We demonstrate experimentally that this perceived information depends critically on a system's network topology. Applying our framework to several real networks, we find that they communicate a large amount of information (having high entropy) and do so efficiently (maintaining low divergence from human expectations). Moreover, we show that such efficient communication arises in networks that are simultaneously heterogeneous, with high-degree hubs, and clustered, with tightly-connected modules – the two defining features of hierarchical organization. Together, these results suggest that many communication networks are constrained by the pressures of information transmission, and that these pressures select for specific structural features.

7.1 INTRODUCTION

Humans receive information in discrete chunks, which transition from one to another – as words in a sentence or notes in a musical progression – to create coherent messages. The networks formed by these chunks (nodes) and transitions (edges) encode the structure of allowed messages, fundamentally governing the ways that we communicate with one another. Although attempts to study the information properties of such transition networks date to the foundation of information theory itself (*603*), with applications to linguistics (*48*, *193*), music theory (*149*), social and information networks (*263*, *570*), the Internet (*405*), and transportation (*571*), fundamental questions concerning the impact of network structure on how humans process information remain unanswered.

The primary difficulty in quantifying the information content of a message is accounting for the human perspective: formally, a message's information content is not inherent, but rather depends crucially on the receiver's expectations (or estimated probabilities) of different symbols and stimuli (*161*, *193*, *603*). Whereas for computers the probabilities of different symbols are often prescribed, human expectations are biased (*309*) and differ from person to person (*193*), with measurable consequences for behavior (*387*) and cognition (*379*). However, recent advances in psychology and neuroscience have shed light on how humans learn and internally estimate the structure of complex probabilistic systems (*180*, *351*, *414*, *419*, *576*, *584*). Given this progress, it is now possible and compelling to build a framework to quantify human information processing and to consider what types of networks support efficient communication.

7.2 HUMANS PERCEIVE INFORMATION BEYOND ENTROPY

We set out to study the amount of information a human perceives when observing a sequence of stimuli. Naturally, one might naively expect a human to perceive roughly the same amount of information as is being produced by a sequence, or its Shannon entropy (*161*, *603*). Here, to motivate our analytic results, we carry out a set of experiments showing that these two quantities – the information perceived by a human and the information produced by a sequence – differ systematically. To experimentally measure perceived information, we employ a paradigm recently developed in statistical learning (*351*, *414*, *419*, *584*), presenting subjects with sequences of stimuli on a screen (Fig. 7.1a) and asking them to respond to each stimulus by pressing the indicated keys on a keyboard (Fig. 7.1b). Although many real communication systems have long-range correlations, the production of information is traditionally modeled as a Markov process (*161*, *603*), or equivalently, a random walk on a (possibly weighted, directed) network (*570*). Therefore, we assign each stimulus to a node in an underlying network, and we stipulate the order of stimuli within a sequence using random walks (Fig. 7.1b; Methods). By measuring subjects' reaction times and error rates, we can infer how much information they perceive: slow reactions or many errors reflect surprising transitions (with high perceived information), while fast reactions or few errors indicate expected transitions (with low perceived information) (*351*, *387*, *419*).

In a random walk, the probability of transitioning from node (or stimulus) i to a neighboring node j is $P_{ij} = 1/k_i$, where k_i is the degree of node i. Thus, the amount of information produced by a single transition $i \to j$ (often referred to as surprisal (*603*)) is given by $-\log P_{ij} = \log k_i$ (Fig. 7.1d) (*161*). Indeed, subjects' behavior is remarkably well-predicted by the information surprisal, with each additional bit of produced information inducing a linear 32 ms increase in reaction times (Fig. 7.1e) and a 0.3% increase in the number of errors (see Sec. 7.8). However, even if we present subjects with networks of constant degree – forcing each transition to produce an identical amount of information – we still discover consistent variations in behavior that are driven by network topology.

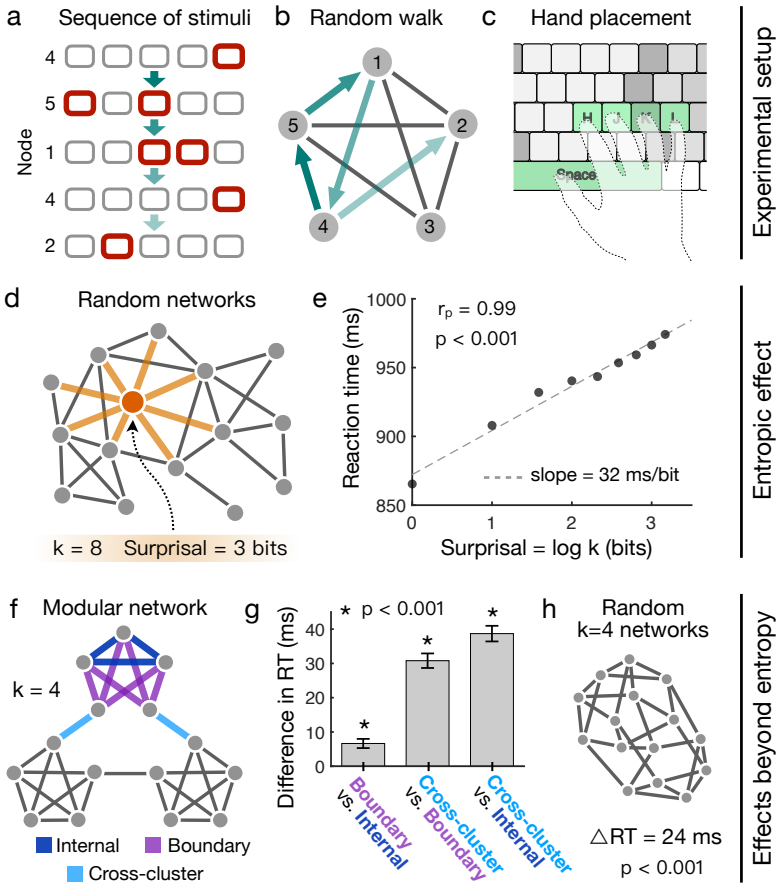


Figure 7.1: Human behavioral experiments reveal the dependence of perceived information on network topology. (*a-c*) Experimental setup for our serial reaction time tasks. (*a*) Subjects are shown sequences of 1500 stimuli, with each stimulus consisting of five squares with one or two highlighted in red. (*b*) The sequential order of stimuli is determined by a random walk on an underlying network. (*c*) In response to each stimulus, subjects press keys on a keyboard corresponding to the highlighted squares. We use both one- and two-button responses because they allow for networks of size up to $N = 15$. To control for the behavioral effects of the different one- and two-button responses, we (i) randomize the assignment of stimuli to nodes for each subject and (ii) regress out behavioral dependencies on individual stimuli (*351*). (*d-e*) Effect of produced information on reaction times, referred to as the entropic effect. (*d*) For each subject, we draw an Erdös-Rényi random network with $N = 15$ nodes and $E = 30$ edges; the information produced by a transition $i \to j$ (or its surprisal) is $\log k_i$, where k_i is the degree of node i. (*e*) Reaction times, averaged over all transitions that begin at nodes of a given degree k, are significantly correlated with the produced information $\log k$ (Pearson correlation coefficient $r_p = 0.99$, $p < 0.001$, $n = 177$ subjects). (*f-h*) Effects of network topology on reaction times after controlling for produced information. (*f*) We control for variations in produced information by focusing on networks of constant degree $k = 4$, such as the modular network, which contains three distinct types of transitions: those deep within clusters (dark blue), those at the boundaries of clusters (purple), and those between clusters (light blue). (*g*) Each type of transition produces reaction times that are distinct from the other two; differences in reaction times and p-values are estimated using mixed effects models ($n = 173$ subjects; see Sec. 7.8.4). (*h*) The difference in reaction times ΔRT between random degree-4 networks and the modular network; the modular network yields consistently faster reactions ($n = 84$ subjects). In addition to the population-level results in panels *e*, *g*, and *h*, we also find significant individual variation in subjects' sensitivity to network topology (see Sec. 7.8.7).

For example, consider the modular network in Fig. 7.1f, which by symmetry only contains three types of transitions. Each transition produces reaction times and error rates that are distinct from the other two (Fig. 7.1g), with transitions between or at the boundaries of clusters generating longer reaction times and more errors (see Sec. 7.8) than those deep within a cluster. In addition to differences in behavior at the level of individual transitions, we also find overall variations between different networks. Specifically, when compared to random networks of constant degree (Fig. 7.1g), the modular network yields significantly faster reactions (and swifter learning rates; see Sec. 7.8.6), indicating a decrease in the average perceived information. Moreover, similar effects have recently been demonstrated across a range of experimental settings (*414*), including networks of varying size and topology (*351*, *362*, *419*); networks with weighted edges (*180*, *445*, *576*); time-varying networks (*419*, *445*); different types of stimuli (*180*, *362*, *445*, *576*, *584*, *667*); and various behavioral and cognitive measures (*180*, *576*, *584*). Together, these results reveal that humans perceive information – beyond the information produced by a sequence – in a manner that depends systematically on network topology.

7.3 QUANTIFYING PERCEIVED INFORMATION: CROSS ENTROPY

The differences between perceived information and produced information can be understood as stemming from the inaccuracy of human expectations. As discussed above, given a transition probability matrix P, a transition $i \to j$ produces $-\log P_{ij}$ bits of information. By contrast, to a person with estimated transition probabilities $\hat{P}$, the same transition will convey $-\log \hat{P}_{ij}$ bits of information.

Although several models have been proposed for how humans estimate transition probabilities (*180*, *414*, *445*, *584*), converging evidence indicates that humans integrate transitions over time (*170*, *242*, *247*, *324*, *419*). Such temporal integration yields expectations that include higher powers of the transition matrix: $\hat{P} = C \sum_{t=0}^{\infty} f(t) P^{t+1}$, where $f(t) \geqslant 0$ is a decreasing function and $C = (\sum_t f(t))^{-1}$ is a normalization constant (we note that $\hat{P}$ is guaranteed to converge if $\sum_t f(t)$ converges). For example, if $f(t) = 1/t!$, then the transition probability estimates $\hat{P}$ are nearly identical to the network communicability (*209*, *242*) (see Sec. 7.8.3). Here, we focus on the specific choice $f(t) = \eta^t$, where $\eta \in (0, 1)$ represents the inaccuracy of a person's expectations (Fig. 7.2a). This model can be derived from a number of different cognitive theories – including the temporal context model of episodic memory (*324*), temporal difference learning and the successor representation in reinforcement learning (*170*, *247*), and the free energy principle from information theory (*419*). Inferring η from each subject's reaction times (Fig. 7.2b; see Methods), we find that 10% of subjects hold exact estimates of the transition structure ($\eta \to 0$; Fig. 7.2a, left), while 21% have expectations that are completely disordered ($\eta \to 1$; Fig. 7.2a, right). Importantly, most subjects have expectations that lie between these two extremes (Fig. 7.2a, center), yielding a decrease in the expected probability of between- versus within-cluster transitions in the modular network. This decrease in expected probability, in turn, gives rise to an increase in perceived information,

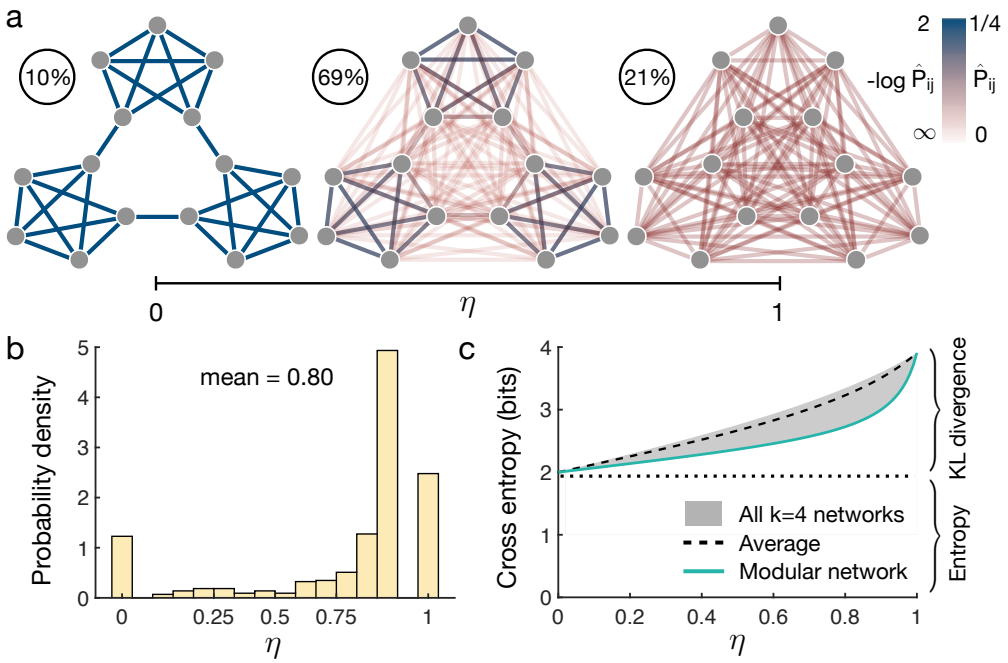


Figure 7.2: Modeling human estimates of transition probabilities. (*a*) Illustration of the internal estimates of the transition probabilities $\hat{P}$ in the modular network. For $\eta \to 0$ (left), the estimates become exact, while for $\eta \to 1$ (right), the estimates become all-to-all, losing any resemblance to the true network. For intermediate η (center), transitions within clusters maintain higher probabilities (and therefore lower surprisal) than transitions between clusters, thereby explaining the differences in reaction times in Fig. 7.1g. Percentages indicate the proportion of subjects, across all tasks, belonging to each category. (*b*) Distribution of the accuracy parameter η estimated from subjects' reaction times (see Sec. 7.8.3); the distribution is over all 518 completed tasks ($n = 434$ subjects). (*c*) Cross entropy $S(P, \hat{P})$ as a function of η for all k-4 networks of size $N = 15$ (shaded region). The modular network (solid line) maintains a lower cross entropy than the average across all k-4 networks (dashed line), thereby explaining the difference in reaction times in Fig. 7.1h.

thereby explaining the observed variations in subjects' reaction times and error rates for different parts of the modular network (Fig. 7.1g).

We are now prepared to study the average perceived information of an entire communication network. Averaging the perceived information of individual transitions over the random walk process, we have $\langle -\log \hat{P}_{ij} \rangle_P = -\sum_{ij} \pi_i P_{ij} \log \hat{P}_{ij}$, where $\boldsymbol{\pi}$ is the stationary distribution of P. Interestingly, this quantity – known as the *cross entropy* $S(P, \hat{P})$ between P and $\hat{P}$ – splits naturally into the entropy $S(P)$, or the average produced information, and the KL divergence $D_{KL}(P||\hat{P})$, or the inefficiency of the observer's expectations:

$$\underbrace{\langle -\log \hat{P}_{ij} \rangle_P}_{S(P,\hat{P})} = \underbrace{\langle -\log P_{ij} \rangle_P}_{S(P)} + \underbrace{\langle -\log \frac{\hat{P}_{ij}}{P_{ij}} \rangle_P}_{D_{KL}(P||\hat{P})}. \tag{7.1}$$

This relationship has a number of immediate consequences, including the fact that the information a human perceives $S(P, \hat{P})$ is lower-bounded by the information that a system produces $S(P)$ (since $D_{KL}(P||\hat{P}) \geqslant 0$). Moreover, inefficiency is minimized when a person's expectations are exact (since $D_{KL}(P||\hat{P}) = 0$ only when $\hat{P} = P$) (*161*). For example, consider the set of degree-4 networks from our human experiments (Fig. 7.1h). While all such networks have identical entropy, their differing topologies induce a range of cross entropies, which vary as a function of η (Fig. 7.2c). Notably, the modular graph displays lower cross entropy than most other degree-4 networks (Fig. 7.2c), thus explaining the observed difference in subjects' behaviors (Fig. 7.1h).

7.4 INFORMATION PROPERTIES OF REAL COMMUNICATION NETWORKS

Using the framework developed above, we are ultimately interested in characterizing the perceived information generated by real communication systems. The networks chosen (Tab. 7.1) have all either evolved or been designed to communicate information through sequences of stimuli (such as words or musical notes) or concepts (such as scientific papers, websites, or social interactions). Strikingly, we find that the networks share two consistent properties: they produce large amounts of information (high entropy; Fig. 7.3a), while at the same time maintaining low inefficiency (low KL divergence; Fig. 7.3b). Specifically, these properties hold relative to completely randomized versions of the networks (Tab. 7.1) , with η set to the average value 0.8 from our human experiments (Fig. 7.2b). Interestingly, different network types exhibit these information properties to varying degrees (Fig. 7.3c). For example, language networks have the highest entropy but also the highest KL divergence, perhaps reflecting the pressure on language to maximize information rate. Meanwhile, music networks are low in both entropy and KL divergence, possibly mirroring their role as a means for entertainment rather than rapid communication.

If we instead compare the communication networks against randomized versions that preserve node degrees (*431*), we find that the entropy is unchanged (Fig. 7.3d), indicating that produced information depends only on the degree distribution. By contrast, even compared to these entropy-preserving networks, the KL divergence of real networks remains low (Fig. 7.3e). We verify that these results largely hold for (i) all values of η, (ii) different models of human expectations $\hat{P}$, and (iii) directed versions of the above networks (Sec. 7.8.8). Moreover, we find that the information properties of communication networks can vary dramatically in time (*184*, *452*), with most networks dynamically evolving (for example, over the course of a musical piece or the growth of a social network) to optimize efficient communication – that is, to maximize entropy and minimize divergence from human expectations (Sec. 7.8.9).

Finally, to demonstrate that efficient communication is not required by all real communication networks, it is important to consider examples where the results in Fig. 7.3 break down. We give two such examples in Sec. 7.8.10, showing that (i) directed citation networks have markedly low entropy and (ii) transitions between words of all parts of speech have relatively high KL divergence. However, if we

Type / Name	N	E	S^{real} (bits)	S^{rand} (bits)	D_{KL}^{real} (bits)	D_{KL}^{rand} (bits)
Language (noun transitions)						
Shakespeare	11,234	97,892	6.15	4.16	1.74	2.17
Homer	3,556	23,608	5.25	3.79	1.75	2.12
Plato	2,271	9,796	4.41	3.19	1.74	2.04
Jane Austen	1,994	12,120	4.92	3.66	1.71	2.10
William Blake	370	781	2.59	2.24	1.64	1.77
Miguel de Cervantes	6,090	43,682	5.55	3.89	1.76	2.14
Walt Whitman	4,791	16,526	4.24	2.89	1.76	2.00
Semantic relationships						
Bible	1,707	9,059	4.31	3.48	1.45	2.07
Les Miserables	77	254	3.25	2.82	0.84	1.65
Edinburgh Thesaurus	7,754	226,518	6.26	5.88	2.07	2.21
Roget Thesaurus	904	3,447	3.19	3.02	1.76	1.99
Glossary terms	60	114	2.32	2.09	1.29	1.55
FOLDOC	13,274	90,736	4.11	3.83	1.72	2.14
ODLIS	1,802	12,378	4.59	3.83	1.70	2.11
World Wide Web						
Google internal	12,354	142,296	6.15	4.56	1.35	2.19
Education	2,622	6,065	3.01	2.36	0.92	1.85
EPA	2,232	6,876	3.34	2.74	1.75	1.95
Indochina	9,638	45,886	3.88	3.33	0.58	2.08
2004 Election blogs	793	13,484	5.78	5.11	1.36	2.01
Spam	3,796	36,404	5.30	4.30	1.66	2.16
WebBase	6,843	16,374	3.48	2.41	1.09	1.87
Citations						
arXiv Hep-Ph	12,711	139,500	5.02	4.49	1.68	2.19
arXiv Hep-Th	7,464	115,932	5.56	4.98	1.64	2.20
Cora	3,991	16,621	3.50	3.14	1.48	2.04
DBLP	240	858	3.30	2.93	1.37	1.88
Social relationships						
Facebook	13,130	75,562	4.22	3.59	1.78	2.11
arXiv Astr-Ph	17,903	196,972	5.39	4.49	1.41	2.19
Adolescent health	2,155	8,970	3.22	3.14	1.78	2.03
Highschool	67	267	3.11	3.07	1.15	1.57
Jazz	198	2,742	5.09	4.81	0.94	1.61
Karate club	34	78	2.58	2.32	1.05	1.40
Music (note transitions)						
Thriller – Michael Jackson	67	446	4.03	3.78	0.76	1.38
Hard Day's Night – Beatles	41	212	3.62	3.42	0.49	1.21
Bohemian Rhapsody – Queen	71	961	5.01	4.77	0.55	0.95
Africa – Toto	39	163	3.41	3.13	0.84	1.29
Sonata No 11 – Mozart	55	354	3.91	3.73	0.83	1.28
Sonata No 23 – Beethoven	69	900	4.86	4.72	0.65	0.96
Nocturne Op 9-2 – Chopin	59	303	3.62	3.42	0.95	1.43
Clavier Fugue 13 – Bach	40	143	3.06	2.92	0.89	1.37
Ballade Op 10-1 – Brahms	69	670	4.42	4.31	0.80	1.18

Table 7.1: Properties of the real communication networks examined in this paper. For each network we list its type and name, number of nodes N and edges E, entropy of the real network S^{real} and after randomizing the edges S^{rand}, and KL divergence of the real network D_{KL}^{real} and after randomization D_{KL}^{rand} with η set to the average value 0.80 from our experiments. S^{rand} and D_{KL}^{rand} are averaged over 100 randomizations. For descriptions of and references for these networks, see Sec. 7.8.14.

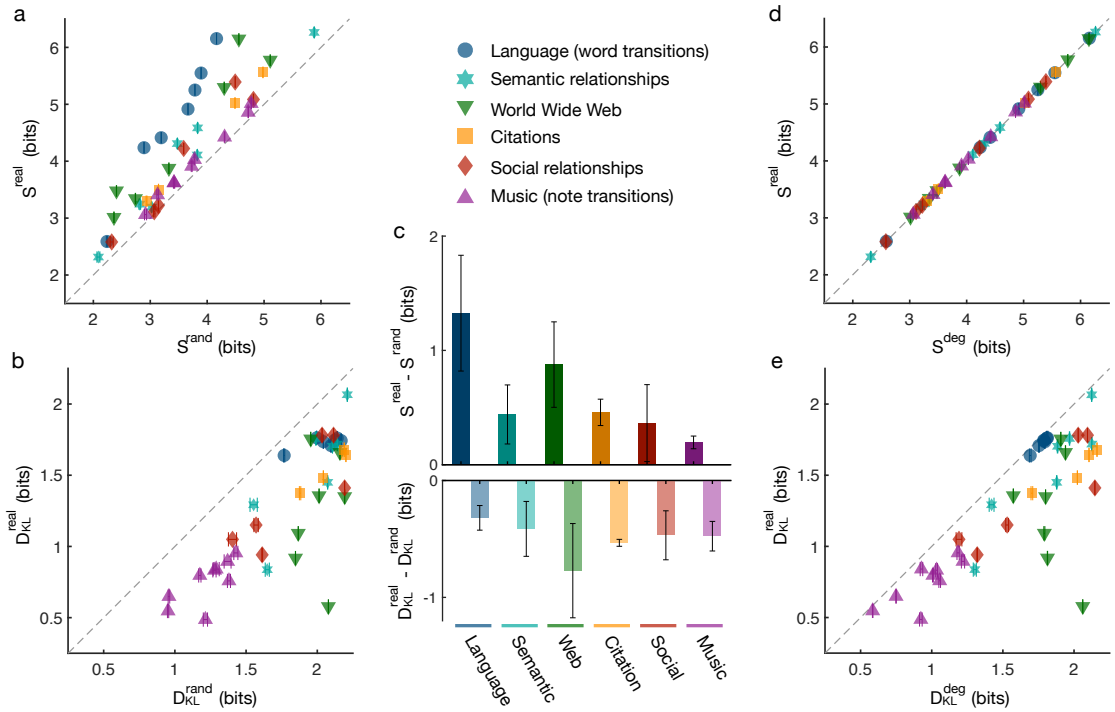


Figure 7.3: The entropy and KL divergence of real communication networks. (*a*) Entropy of fully randomized versions of the networks listed in Tab. 7.1 (S^{rand}) compared with the true values (S^{real}). (*b*) KL divergence of fully randomized versions of the real networks (D_{KL}^{rand}) compared with the true values (D_{KL}^{real}). Human expectations $\hat{P}$ are calculated with η set to the average value 0.80 from our experiments; however, the results remain qualitatively the same across all values of η (Sec. 7.8.8). (*c*) Difference between S^{real} and S^{rand} (top) and difference between D_{KL}^{real} and D_{KL}^{rand} (bottom) for different network types, with error bars indicating standard deviation over networks of each type. (*d*) Entropy of degree-preserving randomized networks (S^{deg}) compared with S^{real}. (*e*) KL divergence of degree-preserving randomized networks (D_{KL}^{deg}) compared with D_{KL}^{real} with fixed $\eta = 0.80$. In panels *a*, *b*, *d*, and *e*, data points and error bars (standard deviations) are estimated from 100 realizations of the randomized networks. All networks are undirected; for examination of directed versions see Sec. 7.8.8.

allow transitions to move both forward and backward along citations (as is typical when traversing scientific literature), then citation networks regain their high entropy (Fig. 7.3a). Similarly, if we focus on "content" words that carry meaning (such as the nouns in Fig. 7.3) rather than "grammatical" words (such as articles, prepositions, and conjunctions) – a common distinction in the study of language networks (*225*, *447*) – then word transitions regain their low KL divergence from human expectations (Figs. 7.3b,e). Thus, even for networks that appear to have high entropy or low KL divergence, studying the context-specific ways that they transmit information to humans often reveals that efficient communication is maintained.

7.5 HIERARCHICALLY MODULAR STRUCTURE

Given the high entropy and low KL divergence displayed by real networks, it is natural to wonder what structural features give rise to these properties. To begin, for undirected networks one can show that $S = \sum_i k_i \log k_i$, demonstrating that the entropy of a network is determined by its degree sequence (Fig. 7.3d) (*117*). It is clear that the entropy grows with increasing node degrees, supporting the intuition that denser networks yield more complex random walks. Moreover, since S is convex in $\mathbf{k}$, the entropy is larger for networks with a small number of high-degree nodes and many low-degree nodes. Interestingly, such heterogeneous structure is observed in human language (*121*), the Internet (*50*), social networks (*477*), and scale-free networks (*50*) (although not all networks with heterogeneous degrees are scale-free (*642*)). To investigate the relationship between a network's entropy and its degree distribution, we derive a number of analytic results in the thermodynamic limit $N \to \infty$ (Sec. 7.8.11). For example, the entropy of an Erdös-Rényi network is given by $S \approx \log\langle k\rangle$ for large average degree $\langle k\rangle$. For scale-free networks with degree exponent γ (Fig. 8.2a), we find that $S = \log\langle k\rangle + \frac{1}{\gamma-2} - \log\frac{\gamma-1}{\gamma-2}$, indicating that $\gamma = 2$ is a critical exponent since the entropy diverges as $\gamma \to 2$. Generating ensembles of Erdös-Rényi and scale-free networks, we numerically verify the logarithmic dependence of S on $\langle k\rangle$ (Fig. 8.2b). Moreover, we find that S increases for decreasing γ (Fig. 8.2c), suggesting that the entropy grows with increasing degree heterogeneity, which we also confirm numerically (Fig. 8.2d). This final result reveals that, after controlling for edge density, the entropy is largest for networks with heavy-tailed degree distributions.

In contrast to the entropy, the KL divergence depends on the expectations of an observer. As these expectations become more accurate (that is, as η decreases), we expect $D_{KL}(P||\hat{P})$ to decrease (as in Fig. 7.2c). But how does the KL divergence depend on network structure? For an undirected network with adjacency matrix G, we can expand in the limit of small η to find that $D_{KL} \approx -\log(1-\eta) - \frac{\eta}{E\ln 2}\sum_i \frac{1}{k_i}\triangle_i$, where $\triangle_i = (G^3)_{ii}/2$ is the number of (possibly weighted) triangles involving node i (Sec. 7.8.12). Therefore, we see that D_{KL} is smaller for networks with a large number of triangles, explaining, for instance, the low KL divergence of the modular network (Figs. 7.1h and 7.2c). Indeed, an abundance of triangles is typically associated with modular structure, a ubiquitous feature of real communication networks, from social and scientific interactions (*250*, *570*) to language (*461*) and the Internet (*207*). To investigate the impact of modularity on the KL divergence, we derive analytic expressions for D_{KL} that hold for all values of η in the thermodynamic limit (Sec. 7.8.12). The KL divergence of an Erdös-Rényi network is given by $D_{KL} = -\log(1-\eta)$. For stochastic block networks with communities of size N_c and a fraction of within-community edges f (Fig. 8.2e), we find that $D_{KL} = -\log\left[1-\eta\left(1-\frac{\langle k\rangle}{N_c}\frac{(1-\eta)f^3}{1-\eta f}\right)\right]$. Generating sets of Erdös-Rényi and stochastic block networks, we confirm the analytic predictions that D_{KL} grows with increasing η (Fig. 8.2f) and decreases for increasing modularity (Fig. 8.2g) and clustering (Fig. 8.2h). Therefore, even after controlling for the inaccuracy

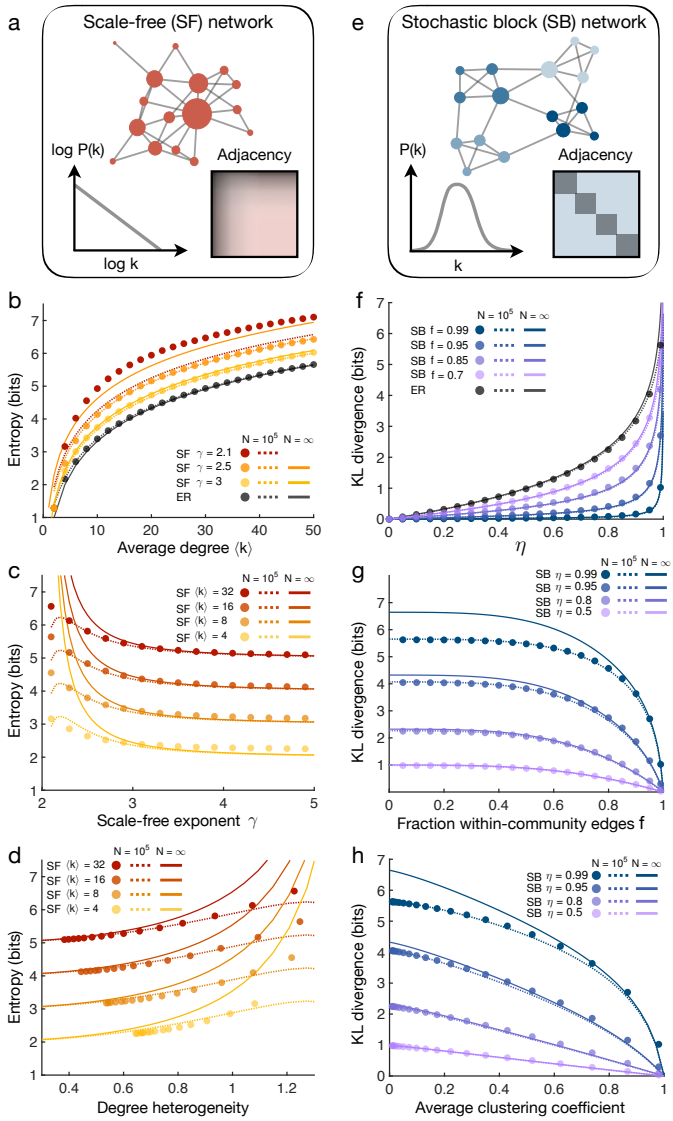


Figure 7.4: The impact of network topology on entropy and KL divergence. (*a*) Scale-free (SF) network, characterized by a power-law degree distribution and the presence of high-degree hub nodes. (*b*) Entropy as a function of the average degree $\langle k \rangle$ for Erdös-Rényi (ER) and SF networks with different scale-free exponents γ. Data points are exact calculations for ER and SF networks generated using the static model (*258*) with size $N = 10^4$. Lines are derived from the expected degree distributions: dashed lines are numerical results for $N = 10^4$ and solid lines are analytic results for $N \to \infty$ (see Sec. 7.8.11 for derivations). Note that the thermodynamic limit for $\gamma = 2.1$ does not appear in the displayed range. (*c*) Entropy as a function of γ for SF networks with fixed $\langle k \rangle$. In the thermodynamic limit (solid lines), the entropy diverges as $\gamma \to 2$, and the analytic results are nearly exact for $\gamma > 3$. (*d*) Entropy as a function of degree heterogeneity $H = \langle |k_i - k_j| \rangle / \langle k \rangle$, where $\langle |k_i - k_j| \rangle$ is the absolute difference in degrees averaged over all pairs of nodes (*410*), for SF networks with fixed $\langle k \rangle$ and variable γ. (*e*) Stochastic block (SB) network, characterized by dense connectivity within communities and sparse connectivity between communities. (*f*) KL divergence as a function of the accuracy parameter η for ER and SB networks with communities of size $N_c = 100$ and different fractions f of within-community edges. Data points are exact calculations for networks with $N = 10^4$ and $\langle k \rangle = 100$, and lines are analytic calculations for $N = 10^4$ (dashed) and $N \to \infty$ (solid; see Sec. 7.8.12 for derivations). (*g*) KL divergence as a function of f for SB networks with fixed η. The analytic results are nearly exact for $\eta < 0.8$. (*h*) KL divergence as a function of the average clustering coefficient for SB networks with fixed η and variable f.

η of human expectations, we find that modular organization serves to decrease the inefficiency of information transmission.

To attain both the high entropy and low KL divergence observed in real communication systems, it appears that networks must be simultaneously heterogeneous and modular, the two defining features of hierarchical organization (*547*). In order to test this hypothesis, we employ a model that combines the heterogeneous degrees of scale-free networks with the modular structure of stochastic block networks (Fig. 7.5a; see Sec. 7.8.13 for an extended description). By adjusting γ and f, we show that these hierarchically modular networks display both a range of entropies (Fig. 7.5b) and KL divergences (Fig. 7.5c). In fact, while scale-free networks do not exhibit the low KL divergence of real communication networks nor do stochastic block networks display their high entropy, we find that hierarchically modular networks can attain both properties (Fig. 7.5d). Taken together, these results indicate that heterogeneity and modularity – precisely the features commonly observed in real communication systems (*50*, *121*, *207*, *250*, *461*, *477*, *547*, *570*) – are both required to achieve high information production and low inefficiency.

7.6 CONCLUSIONS AND OUTLOOK

In this study, we develop tools to quantify the information humans receive from complex networks. We demonstrate experimentally that humans perceive information, beyond the information produced by a sequence, in a way that depends critically on network topology. Moreover, we find that real communication networks support the rapid and efficient transmission of information, and that this efficient communication arises from hierarchical organization. These results raise a number of questions concerning the relationship between human cognition and the structure of communication systems. For example, how have communication networks evolved over time – or perhaps even co-evolved with the brain (*176*) – to facilitate information transmission? Furthermore, how can we design communication systems, from human-technology interfaces (*188*) to classroom lectures (*297*), to optimize efficient communication? The framework presented here provides the mathematical tools to begin answering these questions.

To conclude, we highlight a number of ways that our work can be systematically generalized to analyze more realistic communication systems. First, while we model the production of information as a Markov process (equivalently, a random walk), future work should incorporate the long-range dependencies present in many real communication systems (*112*, *496*). The primary difficulty, however, lies in understanding how humans estimate non-Markov transition structures, with most existing work in statistical learning and artificial grammars focusing on Markov processes (*180*, *250*, *351*, *362*, *414*, *419*, *444*, *576*, *584*, *667*). Second, while we have used tools from information theory to quantify the perceived information of a network (*161*, *603*), these methods do not incorporate the semantic information carried by individual nodes (e.g., words, notes, concepts) (*48*, *193*). Thus, in order to improve our understanding of real-world

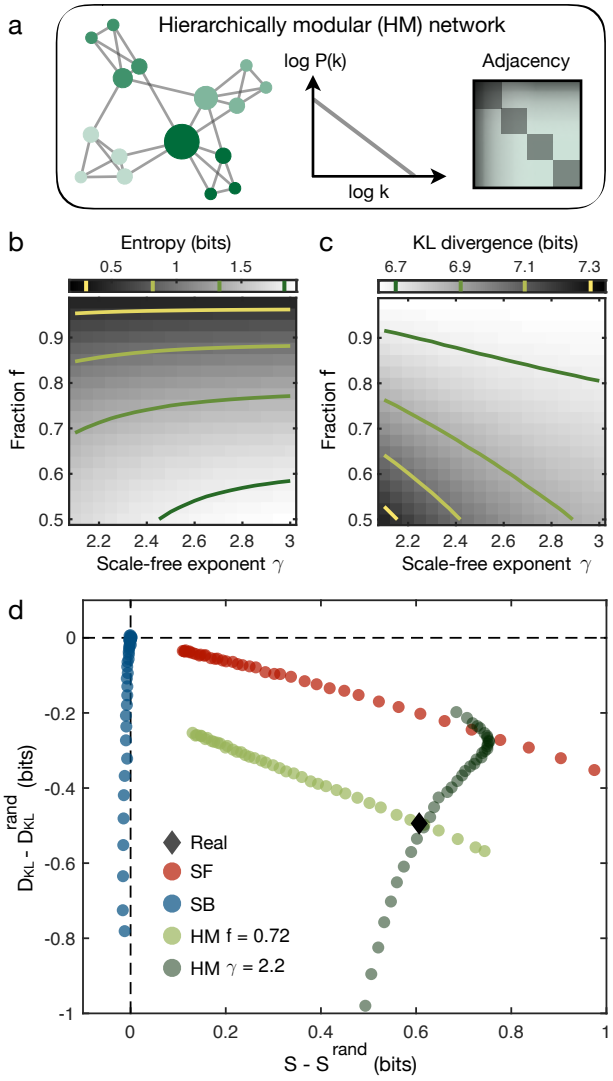


Figure 7.5: Hierarchically modular networks support the efficient communication of information. (*a*) Hierarchically modular (HM) network, characterized by a power-law degree distribution and modular structure (Sec. 7.8.13). (*b*) Entropy as a function of the scale-free exponent γ and the fraction of within-community edges f for HM networks with size $N = 10^4$, average degree $\langle k \rangle = 100$, and community size $N_c = 100$. Solid lines denote networks of equal entropy. (*c*) KL divergence as a function of γ and f for HM networks with the same size and density as panel *b* and η set to the average value 0.80 from our experiments (Fig. 7.2b). Solid lines denote networks of equal KL divergence. (*d*) Average entropies and KL divergences of real and model networks compared to fully randomized versions. Data points are averages over the set of networks in Tab. 7.1, where for each real network we generate SF networks with variable γ (red), SB networks with communities of size $n \approx \sqrt{N}$ and variable f (blue), and HM networks with $n \approx \sqrt{N}$ and variable γ (fixed $f = 0.72$; light green) or variable f (fixed $\gamma = 2.2$; dark green), all with N and E equal to the real network. HM networks with $\gamma = 2.2$ and $f = 0.72$ yield the same average entropy and KL divergence as real communication networks.

communication systems, future progress will require important interdisciplinary efforts from both cognitive scientists (to study how humans estimate non-Markov structures) and information theorists (to quantify semantic information in human contexts).

7.7 METHODS

7.7.1 *Experimental setup*

Subjects performed a self-paced serial reaction time task using a computer screen and keyboard. Each stimulus was presented as a horizontal row of five grey squares; all five squares were shown at all times. The squares corresponded spatially with the keys 'Space', 'H', 'J', 'K', and 'L' (Fig. 7.1c). To indicate a target key or pair of keys for the subject to press, the corresponding squares would become outlined in red (Fig. 7.1a). When subjects pressed the correct key combination, the squares on the screen would immediately display the next stimulus. If an incorrect key or pair of keys was pressed, the message 'Error!' was displayed on the screen below the stimulus and remained until the subject pressed the correct key(s). The order in which stimuli were presented to each subject was determined by a random walk on a network of $N = 15$ nodes. For each subject, one of the 15 key combinations was randomly assigned to each node in the network.

In the first experiment, each subject was assigned an Erdös-Rényi network with $E = 30$ edges. In the second experiment, all subjects responded to sequences of stimuli drawn from the modular network (Fig. 7.1f), which has the same number of nodes and edges. We remark that each node in the modular network is connected to four other nodes, so the entropy of each transition was a constant $-\log \frac{1}{4} = 2$ bits. Some subjects performed both of the first two experiments in back to back stages, with the order of the experiments counterbalanced across subjects. In the third experiment, subjects underwent two stages. In one stage subjects responded to stimuli drawn from the modular network, while in the other stage each subject was assigned a random k-4 network. The order of the two stages was counterbalanced. For each stage of each experiment, subjects responded to sequences of 1500 stimuli.

7.7.2 *Experimental procedures*

All participants provided informed consent in writing and experimental methods were approved by the Institutional Review Board of the University of Pennsylvania. In total, we recruited 363 unique participants to complete our studies on Amazon's Mechanical Turk: 106 completed just the first experiment, 102 completed just the second experiment, 71 completed both the first and second experiments in back-to-back stages, and 84 completed the third experiment. Worker IDs were used to exclude duplicate participants between experiments, and all participants were financially remunerated for their time. In the first two experiments, subjects were paid \$3-\$11 for up to an

estimated 30-60 minutes: $3 per network for up to two networks, $2 per network for correctly responding on at least 90% of the trials, and $1 for completing two stages. In the third experiment, subjects were paid up to $9 for an estimated 60 minutes: $5 for completing the experiment and $2 for correctly responding on at least 90% of the trials on each stage.

7.7.3 *Data analysis*

To make inferences about subjects' internal expectations based upon their reaction times, we excluded all trials in which subjects responded incorrectly. We also excluded reaction times that were implausible, either three standard deviations from a subject's mean reaction time, below 100 ms, or over 3500 ms.

7.7.4 *Measuring the effects of topology on reaction times*

In order to estimate the effects of network topology on subjects' reaction times, one must overcome large inter-subject variability. To do so, we used linear mixed effects models, which have become prominent in human research where many measurements are made for each subject (*582*). Compared with standard linear models, mixed effects models allow for differentiation between effects that are subject-specific and those that are representative of the prototypical individual in our experiments. Here, all models were fit using the `fitlme` function in MATLAB (R2018a), and random effects were chosen as the maximal structure that (i) allowed the model to converge and (ii) did not include effects whose 95% confidence intervals overlapped with zero. In what follows, when referring to our mixed effects models, we employ the standard R notation.

For the first experiment, in order to measure the impact of entropy on reaction times (Fig. 7.1e), we regressed out a number of biomechanical dependencies: (i) variability due to the different button combinations, (ii) the natural quickening of reactions with trial number, and (iii) the change in reaction times between stages. We also regressed out the effects of recency on subjects' reaction times. Specifically, we fit a mixed effects model with the formula '$\text{RT} \sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \,|\, \text{ID})$', where RT is the reaction time, Trial is the trial number (it is common to consider $\log(\text{Trial})$ rather than the trial number itself (*351*, *419*)), Stage is the stage of the experiment, Target is the target button combination, Recency is the number of trials since the last instance of the current stimulus, and ID is each subject's unique ID.

For the second experiment, to measure differences in reaction times between transitions in the modular network (Fig. 7.1g), we fit a mixed effects model of the form '$\text{RT} \sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Trans\_Type} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \,|\, \text{ID})$', where Trans_Type is a dummy variable representing the type of transition (Fig. 7.1g) and the other variables are defined above. The three models for the three different comparisons are summarized in Tabs. 7.3-7.5.

For the third experiment, to measure the difference in reaction times between the modular network and random k-4 networks (Fig. 7.1h), we fit a mixed effects model of the form 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + \text{Graph} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \mid \text{ID})$', where Graph is a dummy variable representing the type of network (either modular or random k-4). This model is summarized in Tab. 7.6.

7.7.5 *Estimating* η *values*

Given a choice for the parameter η, and given a sequence of past nodes $x_1, \ldots, x_{t-1}$, the internal expectation of the next node x_t is predicted to be $\hat{P}_{x_{t-1},x_t}$. We predict subjects' reaction times $r(t)$ using the linear model $\hat{r}(t) = r_0 - r_1 \log \hat{P}_{x_{t-1},x_t}$, where $-\log \hat{P}_{x_{t-1},x_t}$ is the predicted perceived information at time t. Before estimating η, r_0, and r_1, we regress out subjects' biomechanical dependencies using the mixed effects model 'RT $\sim \log(\text{Trial}) * \text{Stage} + \text{Target} + \text{Recency} + (1 + \log(\text{Trial}) * \text{Stage} + \text{Recency} \mid \text{ID})$', where all variables are defined above. Then, to estimate the model parameters that best describe a subject's reactions, we minimize the root-mean-square error (RMSE) with respect to each subject's reaction times. We note that, given a choice for η, the linear parameters r_0 and r_1 can be calculated analytically. Thus, the estimation problem can be restated as a one-dimensional minimization problem; that is, minimizing RMSE with respect to η. To find the global minimum, we began by calculating RMSE along 101 values for η between 0 and 1. Then, starting at the minimum value of this search, we performed gradient descent until the gradient $\frac{\partial \text{RMSE}}{\partial \eta}$ fell below an absolute value of 10^{-6}. The resulting distribution for η over subjects are shown in Fig. 7.2b. For more details, see Sec. 7.8.3.

7.8 SUPPLEMENTARY MATERIAL

In this Supplementary material, we provide extended analysis and discussion to support the results presented in the main text. In Sec. 7.8.1, we clarify the fundamental differences between our work and previous research on human information processing and complex networks. In Sec. 7.8.2, we give a brief introduction to information theory and provide explicit definitions for the quantities discussed in the main text. In Sec. 7.8.3, we introduce existing research studying how humans form expectations about complex transition networks. In Sec. 7.8.4, we present the effects of graph topology on human reaction times measured in our serial response experiments. We begin in Sec. 7.8.4.1 by demonstrating the impact of entropy on reaction times and then proceed to describe effects beyond entropy (Sec. 7.8.4.2, 7.8.4.3). In Sec. 7.8.8, we verify that our conclusions concerning the information properties of real networks hold for (i) various values of η (Sec. 7.8.8.1), (ii) different models of internal representations (Sec. 7.8.8.2), and (iii) directed versions of the real networks (Sec. 7.8.8.3). In Sec. 7.8.11, we derive analytic results for the entropies of various canonical network families. In Sec. 7.8.12, we derive a number of analytic results concerning the KL divergence between random walks and human expectations. In Sec. 7.8.13, we develop a generative model of hierarchically modular networks that combines the heterogeneity of scale-free networks with the community structure of stochastic block networks. Finally, in Sec. 7.8.14, the real networks analyzed in this work are listed and briefly described.

7.8.1 *Previous work*

Our work builds on a long record of research in information theory (*161*, *603*), network science (*11*, *639*), and cognitive science (*147*, *217*, *576*). Here, we clarify the relationships and differences between our work and earlier research in these areas. In particular, we emphasize two main points:

1. In the study of complex networks, traditional definitions of network complexity focus on the structure of a network itself (*11*, *263*, *405*, *570*, *571*, *639*). While characterizing the inherent complexity of a network is a fascinating problem with numerous applications, many complex systems – from language and music to social networks and literature – exist for the sole purpose of communicating information with and between humans. Therefore, to fully understand the structure of these communication networks, one must consider the perspective of a human observer. In this work, we show that this shift in perspective from inherent complexity to perceived complexity can be formally defined using information theory and provides critical insights into the structure of real communication networks.

2. Significant research in cognitive science and statistical learning has studied how humans build internal expectations about the world around them(*147*, *180*, *217*, *351*, *419*, *445*, *480*, *576*, *584*), generating deep insights about human learning

and behavior. Building upon this work, we consider a complimentary problem that has received far less attention: Given a model of human expectations, what types of structures support efficient human communication? The answer to this question may shed light on the organization of real communication systems and help us to design new systems with desirable properties.

7.8.1.1 *Definitions of network entropy*

Information theory has been linked with network science since its inception, when Shannon estimated the entropy rate of the English language by studying a random walk on the network of word transitions in a book(*603*). Since then, information theory has been used extensively to characterize the structure and function of complex networks (*11*, *263*, *405*, *478*, *569–571*, *615*, *639*). Of particular interest are ongoing efforts studying the entropies of random walks on complex networks. For example, the entropies of a number of canonical network families have been derived, including constant-degree networks (*161*) and power-law distributed networks (*263*). Meanwhile, researchers have developed strategies for maximizing the entropy of random walks by tuning the edge weights in a network (*117*, *148*, *183*, *615*), and it is now known that temporal regularities in random walks reveal key aspects of modularity and community structure (*569*, *570*).

Our work extends these efforts by taking into account human expectations. Specifically, we consider the cross entropy (or perceived information) of random walks relative to human expectations, which can be broken down into network entropy (or produced information) and KL divergence (or the inefficiency of human expectations). Importantly, we discover that the entropy and KL divergence characterize distinct aspects of network structure: while entropy is driven by degree heterogeneity, the KL divergence is determined by a network's modular organization. Additionally, we provide a number of novel results concerning network entropy and KL divergence that may be of independent interest. These include analytic approximations for the entropies of networks with Poisson and exponential degree distributions as well as static model networks (see Sec. 7.8.11) and the KL divergences of Erdös-Rényi and stochastic block networks (see Sec. 7.8.12).

7.8.1.2 *Human information processing*

Efforts to relate human cognition to information theory have a rich history, spanning the fields of cognitive science, psychology, and neuroscience. For example, information theory has been used to study linguistics (*48*, *193*), decision-making (*309*, *708*), Bayesian learning (*493*), neural coding (*559*), and vision (*182*). In fact, the relativity of information – the notion that the amount of information conveyed by a message depends not just on the inherent complexity of the message, but also on the expectations of a receiver – was previously studied in linguistics to understand the dependence of meaning in language on context (*193*). To quantify perceived information, however, one requires a mathematical model of human expectations.

Here, we employ recent models from cognitive science and statistical learning to quantitatively study perceived information. In particular, our experimental results build upon a long line of research in cognitive science linking human reaction times to information processing (*330*, *387*) as well as efforts in statistical learning investigating the relationship between human expectations and the network structure of probabilistic transitions (*147*, *180*, *217*, *351*, *360*, *362*, *419*, *445*, *480*, *576*, *584*). Additionally, our analytical results leverage mathematical models of human expectations that have roots in temporal context and temporal difference learning (*247*, *324*) and also appear in reinforcement learning (*170*, *451*) and statistical learning (*419*, *445*, *584*). Using these models of human expectations $\hat{P}$, we are able to quantify the amount of information $\langle -\log \hat{P} \rangle$ that a human perceives when observing a sequence of stimuli.

7.8.2 *Perceived information*

We introduce a specific definition for the information of a sequence of stimuli as perceived by a human observer. We assume that the sequence is generated according to a Markov process with transition probability matrix P. The amount of information produced by a transition from one stimulus i to another stimulus j is $-\log P_{ij}$ (*603*). To quantify the amount of information produced by the entire sequence (per stimulus), one averages this quantity over the Markov process (*161*),

$$\langle -\log P_{ij} \rangle_P = -\sum_i \pi_i \sum_j P_{ij} \log P_{ij}, \tag{7.2}$$

where $\boldsymbol{\pi}$ is the stationary distribution defined by the stationary condition $\boldsymbol{\pi}^\intercal = \boldsymbol{\pi}^\intercal P$. The average quantity in Eq. (9.1) is known as the *entropy rate* of the sequence, although it is often referred to simply as the *entropy*, and it is denoted by $S(P)$.

While the entropy rate quantifies the amount of information produced by a sequence, we are interested in studying the amount of information that a human perceives when observing such a sequence. Consider a human observer with expectations based on an internal estimate of the transition probabilities $\hat{P}$. When observing a transition from one stimulus i to another stimulus j, the observer perceives $-\log \hat{P}_{ij}$ bits of information, which, when averaged over the Markov process, takes the form

$$\langle -\log \hat{P}_{ij} \rangle_P = -\sum_i \pi_i \sum_j P_{ij} \log \hat{P}_{ij}. \tag{7.3}$$

This quantity is the *cross entropy rate* (or simply the *cross entropy*) $S(P, \hat{P})$ between the Markov process P and the observer's expectations $\hat{P}$.

7.8.2.1 *Cross entropy*

If the observer's expectations are exact (that is, if $\hat{P} = P$), then the cross entropy (Eq. (7.3)) reduces to the entropy (Eq. (9.1)); in other words, if the observer correctly

anticipates the frequency of stimuli, then the amount of information they perceive equals the amount of information produced by the sequence itself. However, if the observer's expectations differ from reality (that is, if $\hat{P} \neq P$), then the observer perceives additional information. To see this relationship, we consider the simple identity,

$$S(P, \hat{P}) = S(P) + D_{KL}(P||\hat{P}), \tag{7.4}$$

where $D_{KL}(P||\hat{P})$ is the Kullback-Leibler (KL) divergence between P and $\hat{P}$, defined by

$$D_{KL}(P||\hat{P}) = \langle -\log \frac{\hat{P}_{ij}}{\hat{P}_{ij}} \rangle_P = -\sum_i \pi_i \sum_j P_{ij} \log \frac{\hat{P}_{ij}}{P_{ij}}. \tag{7.5}$$

Gibbs' inequality (*161*) states that $D_{KL}(P||\hat{P}) \geqslant 0$ for all P and $\hat{P}$, and that $D_{KL}(P||\hat{P}) = 0$ only if $\hat{P} = P$. Therefore, we see that the perceived information (or cross entropy) is lower-bounded by the produced information (or entropy).

7.8.2.2 *Random walks on a network*

Every stationary Markov process is equivalent to a random walk on an underlying (possibly weighted, directed) network, where each state is encoded as a node in the network. Specifically, given a transition probability matrix P, one can choose an adjacency matrix G such that

$$P_{ij} = \frac{1}{k_i^{out}} G_{ij}, \tag{7.6}$$

where $k_i^{out} = \sum_j G_{ij}$ is the out-degree of node i. To develop a number of analytic results, we briefly consider the special case of an undirected network. In this case, the out-degree of a node i is referred to simply as its degree k_i. If G is connected, then there exists a unique stationary distribution over nodes, and it is proportional to the degree vector, such that $\boldsymbol{\pi} = \frac{1}{2E}\mathbf{k}$, where $E = \frac{1}{2}\sum_{ij} G_{ij}$ is the number of edges in the network. Therefore, for random walks on a connected, undirected network, we find that the cross entropy can be written as

$$S(P, \hat{P}) = -\frac{1}{2E} \sum_{ij} G_{ij} \log \hat{P}_{ij}, \tag{7.7}$$

reflecting a weighted average of $-\log \hat{P}_{ij}$ over the edges in the network. Moreover, if we further restrict our focus to unweighted networks, then the entropy takes a particularly simple form (*263*):

$$S(P) = \frac{1}{2E} \sum_i k_i \log k_i. \tag{7.8}$$

In this case, it is clear that the entropy of a random walk is uniquely defined by the degree sequence of the network (*161*), a result that is verified numerically for real networks in Fig. 7.2d.

7.8.3 *Human expectations*

When observing sequences of stimuli, humans constantly rely on their internal estimate of the transition structure to anticipate what is coming next (*330*, *343*, *485*, *635*, *711*). Indeed, building expectations about probabilistic relationships allows humans to perform abstract reasoning (*99*), produce language (*227*), develop social intuition (*272*, *667*), and segment streams of stimuli into self-similar parcels (*554*). Moreover, as discussed above, a person's internal expectations, defined by the estimated transition probability matrix $\hat{P}$, determine the amount of information $S(P, \hat{P})$ that they receive from a transition structure defined by P. To study the cross entropy $S(P, \hat{P})$, we require a model $\hat{P} = F(P)$ of how humans internally estimate transition structures in the world around them.

7.8.3.1 *Temporal integration of stimuli*

Models describing how humans learn and estimate transition structures typically stem from Bayesian inference (*272*, *445*, *524*, *658*) or notions of hierarchical learning (*147*, *180*, *444*, *480*). A common thread across many models is that humans relate stimuli that are not directly adjacent in time (*485*, *584*). These non-adjacent relationships have been hypothesized to reflect planning for the future (*170*, *247*), context-dependent memory effects (*324*, *340*), and even errors in optimal Bayesian learning (*419*, *445*). Independent of the underlying mechanisms, the fact that humans relate non-adjacent stimuli results in a common functional form for the expectations $\hat{P}$ where the true transition structure P is integrated over time. Mathematically, this means that $\hat{P}$ includes higher powers of P:

$$\hat{P} = C(f(0)P + f(1)P^2 + \ldots) = C\sum_{t=0}^{\infty} f(t)P^{t+1}, \tag{7.9}$$

with progressively higher powers down-weighted by a decreasing function $f(t) \geqslant 0$, where $C = (\sum_{t=0}^{\infty} f(t))^{-1}$ is a normalization constant. We remark that $\hat{P}$ in Eq. 7.9 is guaranteed to converge as long as the sum $\sum_{t=0}^{\infty} f(t)$ converges.

There exist a number of simple choices for the function $f(t)$. For example, if people's integration of the transition structure drops off as a power law, then we have $f(t) = (t+1)^{-\alpha}$ with power-law exponent $\alpha > 1$. Instead, if the integration drops off with the factorial of t (that is, if $f(t) = 1/t!$), then $\hat{P} = e^{-1}Pe^{P}$, where e^{P} is the matrix exponential. We remark that this model for $\hat{P}$ is nearly equivalent to the communicability of P from graph theory (*209*), which has recently been used to model human expectations (*250*). In Sec. 7.8.8.2 we study the information properties of real networks under these alternative models of human expectations, finding qualitatively equivalent results to those described in the main text.

7.8.3.2 *Exponential model*

Throughout the main text, we focus on a specific model for $\hat{P}$ in which the integration of the transition structure drops off exponentially, such that $f(t) = \eta^t$, where $\eta \in (0, 1)$ is the integration constant. This model is closely related to the successor representation from reinforcement learning (*170*, *451*), which can be derived from temporal context and temporal difference learning (*247*), and can independently be shown to arise from errors in human cognition (*419*). The model takes the following concise analytic form,

$$\begin{aligned}\hat{P} &= \left(\sum_{t=0}^{\infty} \eta^t\right)^{-1} \sum_{t=0}^{\infty} \eta^t P^{t+1} \\ &= (1-\eta)P \sum_{t=0}^{\infty} (\eta P)^t \\ &= (1-\eta)P(I-\eta P)^{-1},\end{aligned} \tag{7.10}$$

where the second equality follows by noticing that $\sum_{t=0}^{\infty} \eta^t = 1/(1-\eta)$ and the third equality follows from the fact that $\sum_{t=0}^{\infty} (\eta P)^t$ converges to $(I-\eta P)^{-1}$ since the spectral radius $\rho(\eta P) = \eta$ is less than one. In the limit $\eta \to 0$, we see that $\hat{P} \to P$, and hence the estimate becomes equivalent to the true transition structure P (Fig. 7.6a). By contrast, in the limit $\eta \to 1$, we find that $\hat{P} \to \mathbf{1}\boldsymbol{\pi}^\intercal$, where $\mathbf{1}$ is the vector of all ones and $\boldsymbol{\pi}$ is the stationary distribution, such that the expectations lose all resemblance to the true structure (Fig. 7.6a). For intermediate values of η, higher-order features of the network, such as communities of densely-connected nodes, maintain much of their probability weight, while some of the fine-scale features, like the edges between communities, fade away (Fig. 7.6a). This strengthening of expectations for transitions within communities relative to transitions between communities is precisely the effect we observe in human reaction times (Fig. 7.1e).

In order to make quantitative predictions for the KL divergence $D_{KL}(P\|\hat{P})$, it is useful to have an estimate for the integration parameter η based on real human data. We estimate η by making predictions for subjects' reaction times and then minimizing the prediction error with respect to η. Given a sequence of nodes $x_1, \ldots, x_{t-1}$, we note that the reaction to the next node x_t is determined by the perceived information of the transition from x_{t-1} to x_t, with expectations calculated at time $t-1$. Formally, this perceived information is given by $-\log \hat{P}_{x_{t-1}, x_t}$, and we make the following linear prediction for the reaction time,

$$\hat{r}(t) = r_0 - r_1 \log \hat{P}_{x_{t-1}, x_t}, \tag{7.11}$$

where the intercept r_0 represents a person's minimum average reaction time (with perfect anticipation of the next stimulus, $\hat{P}_{x_{t-1}, x_t} = 1$) and the slope r_1 quantifies the strength of the relationship between a person's reactions and their perceived information, measured in units of time per bit. Before estimating the model parameters, we first regress out the dependencies of each subject's reaction times on the button

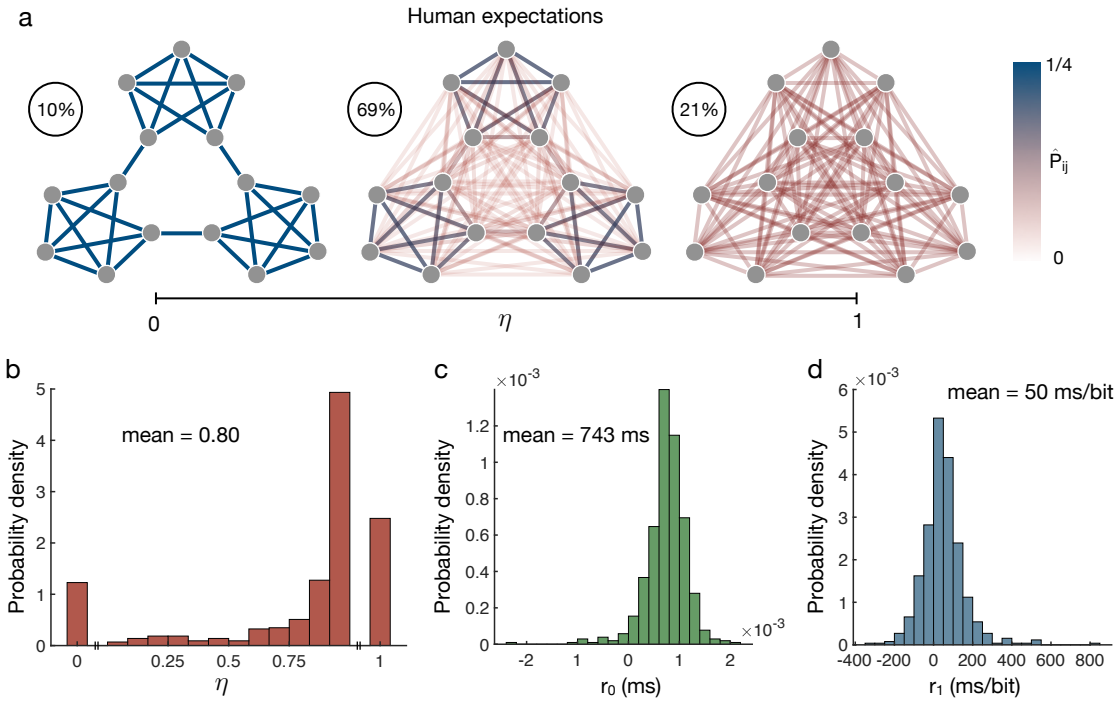


Figure 7.6: Estimated model parameters relating human expectations to reaction times. (*a*) Human expectations $\hat{P}$ for the modular network. For $\eta \to 0$, expectations become exact (left; 10% of subjects), while for $\eta \to 1$, expectations become all-to-all, losing any resemblance to the true structure (right; 21% of subjects). At intermediate values of η, the communities maintain probability weight, while expectations for cross-cluster transitions weaken (center; 69% of subjects). (*b-d*) Distributions of model parameters estimated from subjects' reaction times. Distributions are over all 518 completed sequences. For the integration parameter η (*b*), 53 subjects were best described as having exact representations ($\eta \to 0$) and 107 lacked any notion of the transition structure ($\eta \to 1$), while across all subjects the average value was $\eta = 0.80$. The intercept r_0 is mostly positive (*b*), with an average value of 743 ms. The slope r_1 is also mostly positive (*d*), with an average value of 50 ms/bit.

combinations, trial number, experimental stage, and recency using a mixed effects model of the form '$\mathrm{RT} \sim \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Target} + \mathtt{Recency} + (1 + \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Recency} \,|\, \mathrm{ID})$', where RT is the reaction time, $\mathtt{Trial}$ is the trial number between 1 and 1500 (we found that $\log(\mathtt{Trial})$ was far more predictive of subjects' reaction times than the trial number itself), $\mathtt{Stage}$ is the stage of the experiment (either one or two), $\mathtt{Target}$ is the target button combination, $\mathtt{Recency}$ is the number of trials since the last instance of the current stimulus, and ID is each subject's unique ID. Then, to estimate the parameters η, r_0, and r_1 that best describe a subjects' reaction times, we minimize the RMS error $\sqrt{\frac{1}{T}\sum_t (r(t) - \hat{r}(t))^2}$, where $r(t)$ is the reaction time on trial t after regressing out the above dependencies and T is the number of trials in the experiment. The distributions of the estimated parameters are shown in Fig. 7.6b-d. Among the 518 completed sequences (across 363 unique subjects), 53 were best described as having expectations that exactly matched the transition structure ($\eta \to 0$) and 107 seemed to

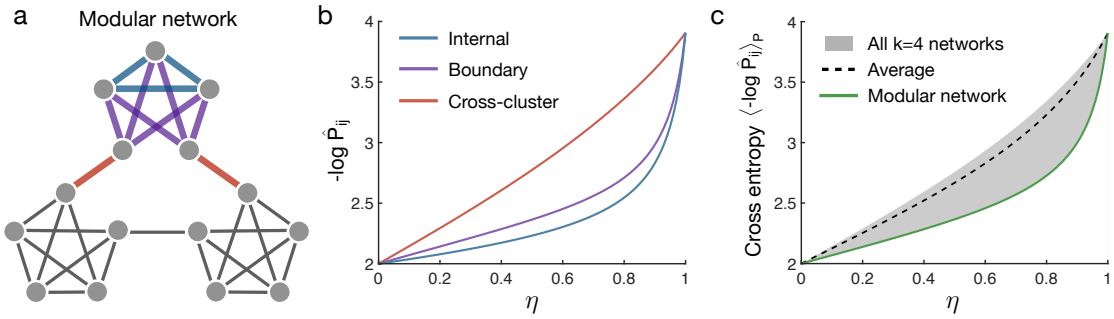


Figure 7.7: Network effects on human reaction times beyond entropy. (*a*) Modular network with three modules of five nodes each. By symmetry the network contains three distinct types of edges: those deep within communities (blue), those at the boundaries of communities (purple), and those between communities (red). (*b*) Perceived information $-\log \hat{P}_{ij}$ for the three edge types as a function of η. Across all values of η, the perceived information is highest for cross-cluster edges, followed by boundary edges, and lowest for internal edges, thus explaining the observed differences in human reaction times (Fig. 7.1e). (*c*) Cross entropy (or network-averaged perceived information) $\langle -\log \hat{P}_{ij} \rangle_P$ as a function of η for the modular network (green) and all k-4 networks (the grey region denotes the range and the dashed line denotes the mean). The modular network maintains nearly the lowest cross entropy among k-4 networks across all values of η, thereby explaining the overall decrease in reaction times in the modular network relative to random k-4 networks (Fig. 7.1f).

lack any notion of the transition structure whatsoever ($\eta \to 1$), with an overall average value of $\eta = 0.80$.

Equipped with the model of human expectations in Eq. (7.10), we can make quantitative predictions for the perceived information of different transition structures. For example, considering the three types of transitions in the modular network (Fig. 7.7a), we find across all values of η that the perceived information $-\log \hat{P}_{ij}$ is highest for transitions between communities, followed by transitions at the boundaries of communities, and lowest for transitions deep within communities (Fig. 7.7b). This prediction precisely matches the variations in reaction times for the different transitions observed in our human experiments (Fig. 7.1e). Furthermore, we find that the average perceived information (or cross entropy) $\langle -\log \hat{P}_{ij} \rangle_P$ is lower in the modular network than almost any other network of the same entropy across all values of η (Fig. 7.7c). This final prediction explains the observed decrease in reaction times in the modular network relative to random entropy-preserving networks (Fig. 7.1f).

7.8.4 *Network effects on reaction times*

In order to directly probe the information that humans perceive, we employ an experimental framework recently developed in statistical learning (*351*, *360*, *362*, *419*, *584*). Specifically, we present human subjects with sequences of stimuli on a computer screen, each stimulus depicting a row of five grey squares with one or two of the squares highlighted in red (Fig. 7.1a, left). In response to each stimulus, subjects are asked to

press one or two computer keys mirroring the highlighted squares (Fig. 7.1a, right). Each of the 15 different stimuli represents a node in an underlying transition network, upon which a random walk stipulates the sequential order of stimuli (Fig. 7.1b). By measuring the speed with which a subject responds to each stimulus, we can infer how much information they are processing – a fast reaction reflects an unsurprising (or uninformative) transition, while a slow reaction reflects a surprising (or informative) transition (*330*, *351*, *387*, *419*, *434*, *635*).

In order to extract the effects of network structure on subjects' reaction times, we use linear mixed effects models, which have become prominent in human research where many measurements are made for each subject (*39*, *582*). To fit our mixed effects models and to estimate the statistical significance of each effect we use the `fitlme` function in MATLAB (R2018a). In what follows, when referring to our mixed effects models, we adopt the standard R notation (*65*).

7.8.4.1 *Entropic effect*

We first investigate the effect of produced information on subjects' reaction times. For undirected and unweighted networks, the produced information (or surprisal) for a single transition from a node i to one of i's neighbors is $\log k_i$, where k_i is the degree of node i. To study a range of surprisal values, we consider completely random networks in which the node degrees are allowed to vary (specifically, we consider random networks with $N = 15$ nodes and $E = 30$ edges). We regress out the dependencies of each subject's reaction times on the button combinations, trial number, experimental stage, and recency using a mixed effects model with the formula '$RT \sim \log(Trial) * Stage + Target + Recency + (1 + \log(Trial) * Stage + Recency \,|\, ID)$', where RT is the reaction time, $Trial$ is the trial number between 1 and 1500, $Stage$ is the stage of the experiment (either one or two), $Target$ is the target button combination, $Recency$ is the number of trials since last observing a node (*41*), and ID is each subject's unique ID. After regressing out these biomechanical dependencies, we find that subjects' average reaction times following nodes of a given degree are accurately predicted by the produced information (Fig. 7.1c), with a Pearson correlation of $r_p = 0.99$ ($p < 0.001$) and a slope of 32 ms/bit.

Additionally, to take into account variations in subjects' reaction times rather than simply studying average reaction times, we employ a mixed effects model of the form '$RT \sim \log(Trial) * Stage + Target + Recency + Surprisal + (1 + \log(Trial) * Stage + Recency + Surprisal \,|\, ID)$', where $Surprisal$ is the logarithm of the degree of the preceding node. The mixed effects model is summarized in Tab. 7.2, reporting a 26 ms increase in reaction times for each additional bit of produced information. We remark that this bit rate is close to that estimated from subjects' average reaction times in random graphs (32 ms/bit; Fig. 7.1c) and is also comparable to the bit rate estimated from our linear prediction of subjects' reaction times in constant-degree graphs (50 ms/bit; Fig. 7.6d).

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1324.8 ± 49.6	26.73	< 0.001	$***$
log(Trial)	-89.6 ± 5.8	-15.41	< 0.001	$***$
Stage	-538.9 ± 54.1	-9.96	< 0.001	$***$
Recency	1.9 ± 0.1	21.63	< 0.001	$***$
Surprisal	26.1 ± 4.1	6.39	< 0.001	$***$
log(Trial):Stage	78.2 ± 6.6	11.91	< 0.001	$***$

Table 7.2: Mixed effects model measuring the effect of produced information on human reaction times. We find a significant 26 ms increase in reaction times ($n = 177$) for each additional bit of produced information, or surprisal (grey). All effects are significant with p-values less than 0.001 ($***$).

7.8.4.2 *Extended cross-cluster effect*

We next investigate reaction time patterns that are driven by perceived information beyond the information produced by a sequence. To experimentally control for the information produced by transitions, we focus on networks of constant degree 4 ($N = 15$ and $E = 30$). Specifically, we consider the modular network shown in Fig. 7.7a, consisting of three communities or clusters comprised of five nodes each. Recent research has shown that people can detect transitions between the clusters (*584*) and that cross-cluster transitions yield increases in reaction times relative to within-cluster transitions (*351*, *419*). These behaviors are surprising in light of the fact that all edges in the network have identical transition probabilities and therefore produce identical amounts of information. Here, we extend these results to include all three of the distinct types of transitions in the modular network (Fig. 7.7a): those deep within communities (internal transitions), those at the boundaries of communities (boundary transitions), and those between communities (cross-cluster transitions).

We use a mixed effects model with the formula '$\mathtt{RT} \sim \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Target} + \mathtt{Recency} + \mathtt{Trans\_Type} + (1 + \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Recency} \mid \mathtt{ID})$', where $\mathtt{Trans\_Type}$ represents the type of transition (either internal, boundary, or cross-cluster). We find a 39 ms increase in reaction times for cross-cluster transitions relative to internal transitions within clusters (Tab. 7.3), a 31 ms increase in reaction times for cross-cluster transitions relative to boundary transitions within clusters (Tab. 7.4), and a 7 ms increase in reaction times for boundary transitions relative to internal transitions within clusters (Tab. 7.5). Notably, this hierarchy of reaction times (Fig. 7.1g) is the same as that predicted by our cross entropy framework (Fig. 7.7b).

7.8.4.3 *Modular effect*

We finally investigate the effects of perceived information averaged over all transitions in a network, defined by the cross entropy in Eq. (7.3). To do so, we compare reaction times in the modular network with reaction times in random k-4 networks. We remark

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1365.6 ± 46.8	29.15	< 0.001	$***$
log(Trial)	-86.9 ± 5.2	-16.75	< 0.001	$***$
Stage	-549.2 ± 52.9	-10.38	< 0.001	$***$
Recency	1.5 ± 0.1	18.40	< 0.001	$***$
Trans_Type	38.7 ± 2.3	16.99	< 0.001	$***$
log(Trial):Stage	63.5 ± 5.8	11.01	< 0.001	$***$

Table 7.3: Mixed effects model measuring the difference in reaction times between internal and cross-cluster transitions. We find a significant 39 ms increase in reaction times ($n = 173$) for cross-cluster transitions relative to internal transitions within communities (grey). All effects are significant with p-values less than 0.001 ($***$).

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1349.3 ± 45.8	29.48	< 0.001	$***$
log(Trial)	-86.0 ± 5.2	-16.39	< 0.001	$***$
Stage	-495.41 ± 49.6	-9.98	< 0.001	$***$
Recency	1.6 ± 0.1	23.28	< 0.001	$***$
Trans_Type	30.8 ± 2.1	14.50	< 0.001	$***$
log(Trial):Stage	62.1 ± 5.8	10.76	< 0.001	$***$

Table 7.4: Mixed effects model measuring the difference in reaction times between boundary and cross-cluster transitions. We find a significant 31 ms increase in reaction times ($n = 173$) for cross-cluster transitions relative to boundary transitions within communities (grey). All effects are significant with p-values less than 0.001 ($***$).

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1333.3 ± 44.3	30.13	< 0.001	$***$
log(Trial)	-84.0 ± 4.9	-17.11	< 0.001	$***$
Stage	-464.8 ± 47.2	-9.84	< 0.001	$***$
Recency	1.5 ± 0.1	24.55	< 0.001	$***$
Trans_Type	6.6 ± 1.3	4.96	< 0.001	$***$
log(Trial):Stage	60.0 ± 5.4	11.12	< 0.001	$***$

Table 7.5: Mixed effects model measuring the difference in reaction times between internal and boundary transitions within clusters. We find a significant 7 ms increase in reaction times ($n = 173$) for boundary transitions relative to internal transitions within communities (grey). All effects are significant with p-values less than 0.001 ($***$).

that the entropy (defined in Eq. (9.1)) is identical across all graphs considered. We use a mixed effects model of the form '$\texttt{RT} \sim \log(\texttt{Trial}) * \texttt{Stage} + \texttt{Target} + \texttt{Recency} +$

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1195.0 ± 48.8	24.49	< 0.001	$***$
log(Trial)	-71.9 ± 4.9	-14.61	< 0.001	$***$
Stage	-405.3 ± 36.9	-10.98	< 0.001	$***$
Recency	1.7 ± 0.1	19.65	< 0.001	$***$
Network_Type	23.5 ± 6.9	3.39	< 0.001	$***$
log(Trial):Stage	49.0 ± 5.1	9.61	< 0.001	$***$

Table 7.6: Mixed effects model measuring the difference in reaction times between the modular network and random k-4 networks. We find a significant 24 ms increase in reaction times ($n = 84$) for random k-4 networks (that is, networks of equal entropy) relative to the modular network (grey). All effects are significant with p-values less than 0.001 ($***$).

$\mathtt{Network\_Type} + (1 + \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Recency} \,|\, \mathrm{ID})$', where $\mathtt{Network\_Type}$ represents the type of network (either modular of random k-4). The estimated mixed effects model is summarized in Tab. 7.6, reporting a 24 ms increase in reaction times for random degree-preserving networks relative to the modular network. Notably, this effect is predicted by our cross entropy framework (Fig. 7.7c). Moreover, this result provides direct evidence that, even after controlling for the entropy of a network, modular structure reduces the total amount of information that humans perceive when observing a sequence of stimuli.

7.8.5 *Network effects on errors*

In addition to measuring the effects of network structure on subjects' reaction times, we can also investigate variations in subjects' error rates. Here, we study the same entropic, extended cross-cluster, and modular effects as in Sec. 7.8.4 above, but on error rates instead of reaction times.

7.8.5.1 *Entropic effect*

We first investigate the effect of produced information (or surprisal) $\log k_i$ on subjects' error rates. Specifically, we consider the same random networks as in Sec. 7.8.4.1. To measure the effect of produced information on error rates, we estimate a mixed effects model of the form '$\mathtt{Error} \sim \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Target} + \mathtt{Recency} + \mathtt{Surprisal} + (1 + \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Recency} + \mathtt{Surprisal} \,|\, \mathrm{ID})$', where $\mathtt{Error}$ equals one for error trials and zero for correct trials, and the other variables have been defined previously. The estimated model is summarized in Tab. 7.7, with a significant 0.3% increase in errors for each additional bit of produced information.

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.078 ± 0.009	8.73	< 0.001	$***$
log(Trial)	-0.007 ± 0.001	-5.57	< 0.001	$***$
Stage	-0.037 ± 0.011	-3.30	< 0.001	$***$
Recency	0.001 ± 0.000	14.01	< 0.001	$***$
Surprisal	0.003 ± 0.001	2.74	0.006	$**$
log(Trial):Stage	0.005 ± 0.002	3.34	< 0.001	$***$

Table 7.7: Mixed effects model measuring the effect of produced information on error rates. We find a significant 0.3% increase in errors ($n = 177$) for each additional bit of produced information, or surprisal (grey). The significance column indicates p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$).

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.068 ± 0.010	7.04	< 0.001	$***$
log(Trial)	-0.005 ± 0.002	-3.30	< 0.001	$***$
Stage	-0.035 ± 0.013	-2.62	0.009	$**$
Recency	0.000 ± 0.000	9.73	< 0.001	$***$
Trans_Type	0.009 ± 0.002	5.69	< 0.001	$***$
log(Trial):Stage	0.006 ± 0.002	3.03	0.002	$**$

Table 7.8: Mixed effects model measuring the difference in error rates between internal and cross-cluster transitions. We find a significant 0.9% increase in errors ($n = 173$) for cross-cluster transitions relative to internal transitions within communities (grey). The significance column indicates p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$).

7.8.5.2 *Extended cross-cluster effect*

We next study variations in error rates that are driven by perceived information after controlling for information produced by a sequence. Considering once again the modular network in Fig. 7.8a, we measure the differences in error rates between the different types of transitions. We use a mixed effects model of the form '$\mathtt{Error} \sim \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Target} + \mathtt{Recency} + \mathtt{Trans\_Type} + (1 + \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Recency} \,|\, \mathrm{ID})$', where $\mathtt{Trans\_Type}$ denotes the type of transition (either internal, boundary, or cross-cluster). We find a significant 0.9% increase in errors for cross-cluster transitions relative to internal transitions within clusters (Tab. 7.8), a significant 0.6% increase in errors for cross-cluster transitions relative to boundary transitions within clusters (Tab. 7.9), and a significant 0.2% increase in errors for boundary transitions relative to internal transitions within clusters (Tab. 7.10). Notably, we find the same hierarchy of effects on error rates (Fig. 7.8) as for reaction times (Fig. 7.1g) and as predicted by our cross entropy framework (Fig. 7.7b).

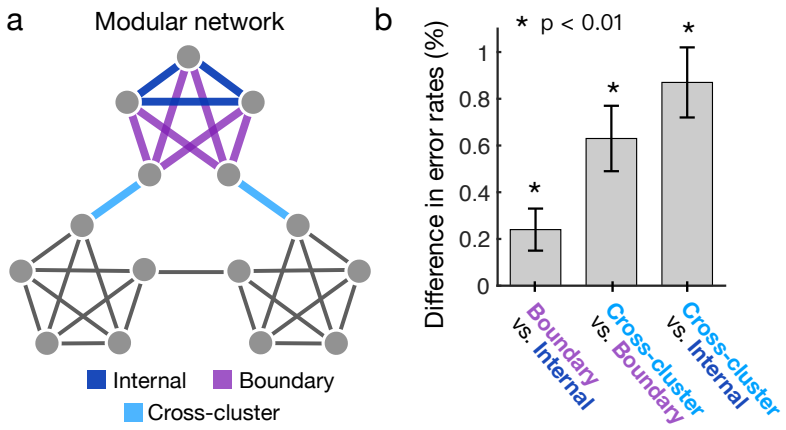


Figure 7.8: Effects of modular topology on error rates. (*a*) Modular network with three types of edges: internal edges within communities (dark blue), boundary edges within communities (purple), and cross-cluster edges between communities (light blue). (*b*) Differences in error rates between the different types of transitions; we find significant differences in error rates between all three types of transitions ($n = 173$ subjects).

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.075 ± 0.008	8.91	< 0.001	$***$
log(Trial)	-0.006 ± 0.001	4.68	< 0.001	$***$
Stage	-0.044 ± 0.013	-3.53	< 0.001	$***$
Recency	0.000 ± 0.000	12.82	< 0.001	$***$
Trans_Type	0.006 ± 0.001	4.38	< 0.001	$***$
log(Trial):Stage	0.008 ± 0.002	3.88	< 0.001	$***$

Table 7.9: Mixed effects model measuring the difference in error rates between boundary and cross-cluster transitions. We find a significant 0.6% increase in errors ($n = 173$) for cross-cluster transitions relative to boundary transitions within communities (grey). All effects are significant with p-values less than 0.001 ($***$).

7.8.5.3 *Modular effect*

Finally, we investigate the effect of the average perceived information (or cross entropy) of a network, while still controlling for produced information. Specifically, we compare subjects' reaction times in the modular network with reaction times in random k-4 networks, noting that the average produced information (or entropy) is identical across all graphs considered. We employ a mixed effects model of the form '$\mathrm{Error} \sim \log(\mathrm{Trial}) * \mathrm{Stage} + \mathrm{Target} + \mathrm{Recency} + \mathrm{Network\_Type} + (1 + \log(\mathrm{Trial}) * \mathrm{Stage} + \mathrm{Recency} \,|\, \mathrm{ID})$', where Network_Type indicates the type of network (either modular of random k-4). Although we find a 0.2% increase in errors for random k-4 networks relative to the modular network, this difference is not significant (Tab. 7.11).

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.070 ± 0.008	8.63	< 0.001	$***$
log(Trial)	-0.006 ± 0.001	-4.11	< 0.001	$***$
Stage	-0.034 ± 0.011	-3.00	0.003	$**$
Recency	0.000 ± 0.000	13.81	< 0.001	$***$
Trans_Type	0.002 ± 0.001	2.66	0.008	$**$
log(Trial):Stage	0.006 ± 0.002	3.45	< 0.001	$***$

Table 7.10: Mixed effects model measuring the difference in error rates between internal and boundary transitions within clusters. We find a significant 0.2% increase in errors ($n = 173$) for boundary transitions relative to internal transitions within communities (grey). The significance column indicates p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$).

Effect	Estimate	t-value	Pr(>|t|)	Significance
(Intercept)	0.038 ± 0.010	3.83	< 0.001	$***$
log(Trial)	-0.004 ± 0.001	-2.49	0.013	$*$
Stage	-0.034 ± 0.011	-2.96	0.003	$**$
Recency	0.000 ± 0.000	9.98	< 0.001	$***$
Network_Type	0.002 ± 0.002	0.96	0.329	
log(Trial):Stage	0.005 ± 0.002	2.82	0.005	$**$

Table 7.11: Mixed effects model measuring the difference in error rates between the modular network and random k-4 networks. We do not find a significant difference in error rates ($n = 84$) between the modular network and random k-4 networks (grey). The significance column indicates p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$).

7.8.6 *Modular effect on learning rate*

In the previous two sections, we investigated the effects of network structure on human reaction times and error rates, without considering the learning dynamics. Here we study the effect of network structure on learning rate, or how quickly subjects' reaction times decrease for a given increase in the number of trials. Specifically, we seek to determine which type of network is *faster* to learn: the modular network (Fig. 7.7a) or the random k-4 networks (Fig. 7.1h). To do so, we estimate a mixed effects model of the form '$\mathrm{RT} \sim \log(\mathtt{Trial}) * \mathtt{Stage} + \log(\mathtt{Trial}) * \mathtt{Network\_Type} + \mathtt{Target} + \mathtt{Recency} + (1 + \log(\mathtt{Trial}) * \mathtt{Stage} + \mathtt{Recency} \,|\, \mathrm{ID})$'. We note that the only difference between this model and that used in Sec. 7.8.4.3 is the interaction term between $\log(\mathtt{Trial})$ and $\mathtt{Network\_Type}$, which tells us how the network type impacts the effect of $\log(\mathtt{Trial})$ on reaction times (or the learning rate). The estimated model is summarized in Tab. 7.12, reporting a significant 9 ms increase in reaction times for each e-fold increase in $\mathtt{Trial}$ for the random k-4 networks relative to the modular network. Intuitively, this

Effect	Estimate (ms)	t-value	Pr(>|t|)	Significance
(Intercept)	1222.6 ± 50.9	24.00	< 0.001	$***$
log(Trial)	-76.0 ± 5.3	-14.21	< 0.001	$***$
Stage	-401.1 ± 36.6	-10.95	< 0.001	$***$
Recency	1.7 ± 0.1	19.65	< 0.001	$***$
Network_Type	-35.9 ± 30.9	-1.16	0.245	
log(Trial):Stage	48.4 ± 5.1	9.58	< 0.001	$***$
log(Trial):Network_Type	8.8 ± 4.4	1.98	0.048	$*$

Table 7.12: Mixed effects model measuring the difference in learning rates between the modular network and random k-4 networks. For each e-fold increase in the number of trials, we find a significant 9 ms increase in reaction times ($n = 84$) for random k-4 networks relative to the modular network (grey). The significance column indicates p-values less than 0.001 ($***$), less than 0.01 ($**$), and less than 0.05 ($*$).

means that the learning rate is faster (that is, reaction times decrease more for each increase in `Trial`) for the modular network than for the k-4 networks.

7.8.7 *Individual differences in network effects*

In the previous three sections, we have discussed the *fixed* effects of network structure on human behavior, which do not vary from person to person. However, for each network effect, we also find a significant amount of variation between individuals. Specifically, for each of the reaction time effects in Sec. 7.8.4 and error rate effects in Sec. 7.8, we fit a mixed effects model that includes a *random* (or *mixed*) effect term that differs for each subject. In this way, we are able to estimate the effect size for each participant in our experiments. For all of the reaction time effects (Fig. 7.9a-e) and all of the error rate effects (Fig. 7.9f-j), we find a significant standard deviation in the distribution of network effects ($p < 0.05$), indicating that each network effect exhibits significant inter-subject variability. Moreover, we find that many of the network effects on reaction times and error rates are significantly correlated across subjects (Fig. 7.10), indicating that they are likely to be driven by common underlying mechanisms.

To understand what might be driving these individual differences in behavior, it helps to recall our linear predictions of subjects' reaction times $\hat{r}(t) = r_0 - r_1 \log \hat{P}_{x_{t-1},x_t}$, where $\hat{r}(t)$ is the predicted reaction time on trial t and $\hat{P}_{x_{t-1},x_t}$ is the model of human transition probability estimates, where x_{t-1} and x_t are the stimuli on trials $t-1$ and t (see Sec. 7.8.3.2). The predictions contain three parameters, which are estimated separately for each subject: the inaccuracy parameter η, which is included in $\hat{P}$ (Fig. 7.6b), the intercept r_0 (Fig. 7.6c), and the slope r_1 (Fig. 7.6d). Among these three parameters, the inaccuracy η has drawn the most attention in the literature, having been shown to correlate with working memory performance (*419*), drive differences

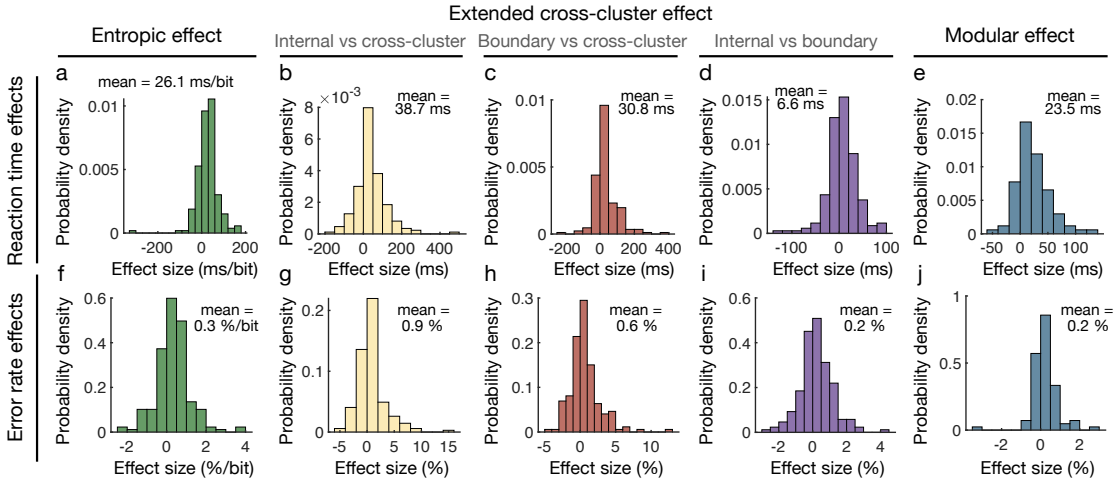


Figure 7.9: Distributions of network effects over individual subjects. (*a-e*) Distributions over subjects of the different reaction time effects: the entropic effect ($n = 177$), or the increase in reaction times for increasing produced information (*a*); the extended cross-cluster effects ($n = 173$), or the difference in reaction times between internal and cross-cluster transitions (*b*), between boundary and cross-cluster transitions (*c*), and between internal and boundary transitions (*d*) in the modular graph; and the modular effect ($n = 84$), or the difference in reaction times between the modular network and random k-4 networks (*e*). (*f-j*) Distributions over subjects of the different effects on error rates: the entropic effect (*f*), the extended cross-cluster effects (*g-i*), and the modular effect (*j*).

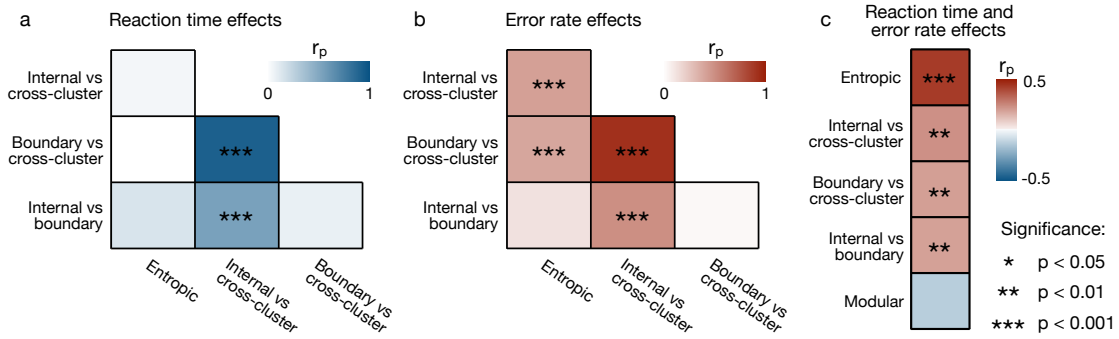


Figure 7.10: Correlations between different network effects across subjects. (*a*) Pearson correlations between the entropic and extended cross-cluster effects on reaction times. (*b*) Pearson correlations between the entropic and extended cross-cluster effects on error rates. In *a* and *b*, the modular effects on reaction times and error rates are not shown because they were measured in a different population of subjects. (*c*) For each network effect, we show the Pearson correlation between the corresponding reaction time effect and error rate effect. Statistically significant correlations are indicated by p-values less than 0.001 (***), less than 0.01 (**), and less than 0.05 (*).

in behaviors in reinforcement learning tasks (*246*), and determine the time-scale of episodic memories in the temporal context model (*247*).

Here, we consider the possible role of η in driving the individual differences in behaviors observed in Fig. 7.9. We first note that we should not expect a monotonic

relationship between η and any of the extended cross cluster effects (Fig. 7.9b-d,g-i) or the modular effects (Fig. 7.9e,j). Indeed, all of these effects disappear in both the high- and low-η limits (Fig. 7.7b,c); for low η, humans have exact representations of the transition network and there will be no difference in the estimated probabilities of different transitions in the modular network or any other k-4 network, while for high η, human estimates of the transition probabilities become completely disordered and, yet again, there is no difference in the estimated transition probabilities. However, for random networks with non-uniform degrees (Fig. 7.1d), as η increases the estimate $\hat{P}$ of the transition network will become less accurate, and therefore the entropic effect (Fig. 7.1e) should become weaker. Indeed, we find a significant negative correlation between η and the entropic effect on reaction times (Spearman correlation $r_s = -0.25$; $p < 0.001$); we note that we use the Spearman correlation coefficient because η is far from normally distributed (Fig. 7.6b). Together, these results demonstrate that there are individual differences in sensitivity to network structure (Fig. 7.9), and that these differences may be related to variations in the accuracy of people's estimates of transition networks.

7.8.8 *Real networks*

In the main text, we show that real networks exhibit two consistent information properties: they have high entropy and low KL divergence from human expectations. When calculating the KL divergence, we use the model $\hat{P}$ defined in Eq. (7.10) with η set to the average value from our human experiments (Fig. 7.6b). Additionally, in order to draw on our analytical results (see Secs. 7.8.11 and 7.8.12), we focused on undirected versions of the real networks. Here, we show that the central conclusions in the main text concerning the information properties of real networks are robust to variations in these choices. Specifically, we verify that the KL divergence of real networks remains low for different values of η and different models for $\hat{P}$ altogether, and we confirm that the entropy remains high and the KL divergence remains low for directed versions of the real networks.

7.8.8.1 *Varying* η

We first investigate how the KL divergence varies as a function of the inaccuracy parameter η. To recall, the KL divergence, defined in Eq. (7.5), represents the inefficiency due to a person's expectations $\hat{P}$. We consider the model of expectations used in the main text, $\hat{P} = (1-\eta)P(I-\eta P)^{-1}$, while varying the parameter η between zero and one. We find that all of the real networks considered maintain a lower KL divergence than fully randomized versions of the networks across all values of η (Fig. 7.11a). In the limit $\eta \to 0$, the KL divergence of both real and randomized networks tends toward zero (Fig. 7.11a), as expected. As η increases, the difference in efficiency between the real and fully randomized networks grows (Fig. 7.11b). We also generate randomized versions of the real networks that maintain identical entropies by preserving the degree distribution. Even when compared against random networks with the same entropy,

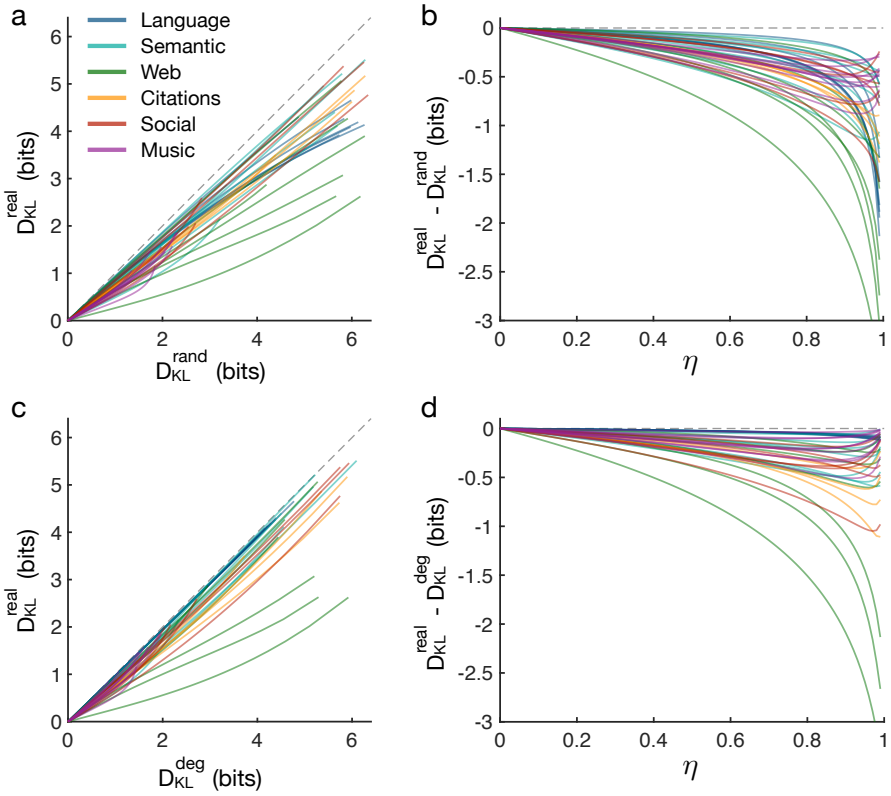


Figure 7.11: KL divergence of real networks for different values of η. (*a*) KL divergence of fully randomized versions of the real networks listed in Tab. 7.13 (D_{KL}^{rand}) compared with the true value (D_{KL}^{real}) as η varies from zero to one. Every real networks maintains lower KL divergence than the corresponding randomized network across all values of η. (*b*) Difference between the KL divergence of real and fully randomized networks as a function of η. (*c*) KL divergence of degree-preserving randomized versions of the real networks (D_{KL}^{deg}) compared with D_{KL}^{real} as η varies from zero to one. The real networks display lower KL divergence than the degree-preserving randomized versions across all values of η. (*d*) Difference between the KL divergence of real and degree-preserving randomized networks as a function of η. All networks are undirected, and each line is calculated using one randomization of the corresponding real network.

all of the real networks attain lower KL divergence across all values of η (Fig. 7.11c). Just as for the fully randomized networks, the difference in efficiency between real and entropy-preserving random networks grows as η increases (Fig. 7.11d). These results confirm that our conclusions in the main text are robust to variations in the inaccuracy parameter η.

7.8.8.2 *Different internal representations*

Here, we study the KL divergence for different models of the human expectations $\hat{P}$. First, we consider the power-law model, defined by Eq. (7.9) with integration function $f(t) = (t+1)^{-\alpha}$, where $\alpha \in (1, \infty)$ is the single parameter. Varying α between 1 and 10,

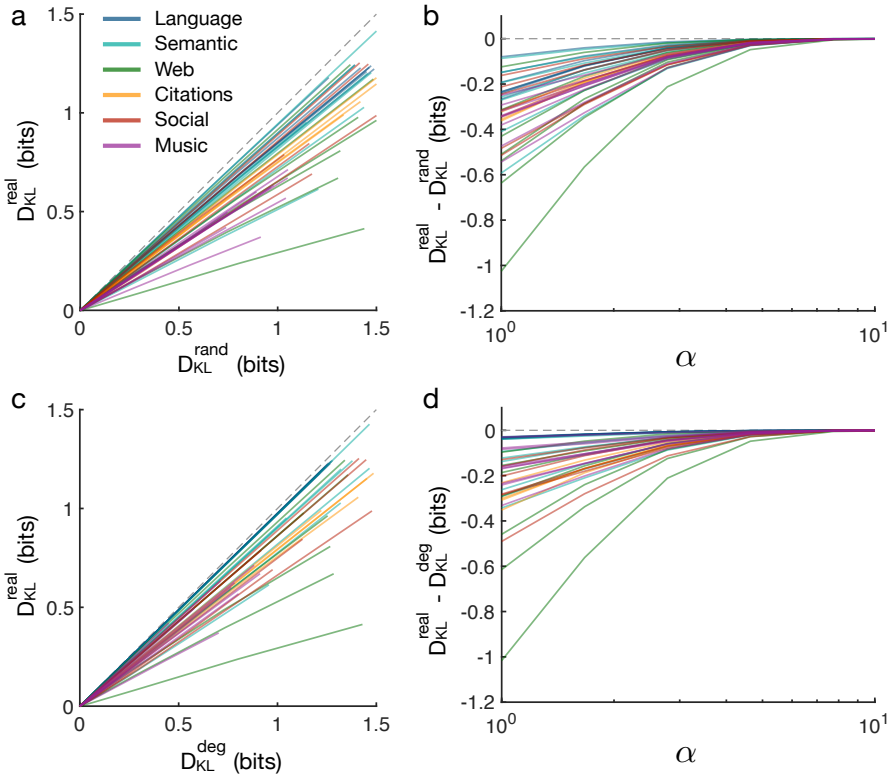


Figure 7.12: KL divergence of real networks under the power-law model of human expectations. (*a*) KL divergence of fully randomized versions of the real networks listed in Tab. 7.13 (D_{KL}^{rand}) compared with the true value (D_{KL}^{real}). Expectations $\hat{P}$ are defined as in Eq. (7.9) with $f(t) = (t+1)^{-\alpha}$, and we allow α to vary between 1 and 10. The real networks maintain lower KL divergence than the randomized network across all values of α. (*b*) Difference between the KL divergence of real and fully randomized networks as a function of α. (*c*) KL divergence of degree-preserving randomized versions of the real networks (D_{KL}^{deg}) compared with D_{KL}^{real} as α varies from 1 to 10. The real networks display lower KL divergence than the degree-preserving randomized versions across all values of α. (*d*) Difference between the KL divergence of real and degree-preserving randomized networks as a function of α. All networks are undirected, and each line is calculated using one randomization of the corresponding real network.

we find that all of the real networks display lower KL divergence than fully randomized versions for all values of α (Fig. 7.12a). Moreover, this difference in efficiency grows as α decreases (Fig. 7.12b); that is, the difference in KL divergence increases as the expectations $\hat{P}$ integrate over longer time scales, which is analogous to η increasing. Even when compared with random versions that preserve the entropy, the real networks still exhibit lower KL divergence across all values of α (Fig. 7.12c,d).

Second, we consider the factorial model for $\hat{P}$, defined by Eq. (7.9) with integration function $f(t) = 1/t!$. As discussed in Sec. 7.8.3.1, this model takes the analytic form $\hat{P} = e^{-1}Pe^{P}$, where e^{P} is the matrix exponential, which is closely related to the communicability of P (*209*, *250*). Calculating the KL divergence, we find qualitatively the same results as for the previous two models. Namely, when compared against

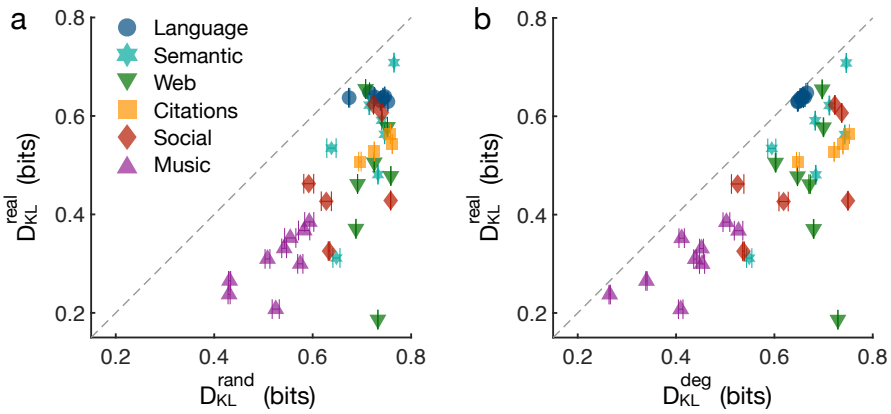


Figure 7.13: KL divergence of real networks under the factorial model of human expectations. (*a*) KL divergence of fully randomized versions of the real networks listed in Tab. 7.13 (D_{KL}^{rand}) compared with the exact value (D_{KL}^{real}). Expectations $\hat{P}$ are defined as in Eq. (7.9) with $f(t) = 1/t!$. (*b*) KL divergence of degree-preserving randomized versions of the real networks (D_{KL}^{deg}) compared with D_{KL}^{real}. In both cases, the real networks maintain lower KL divergence than the randomized versions. Data points and error bars (standard deviations) are estimated from 10 realizations of the randomized networks.

both fully randomized and entropy-preserving (i.e., degree-preserving) randomized versions, all of the real networks studied maintain a lower KL divergence (Fig. 7.13). Taken together, the results of this and the previous subsections indicate that the low KL divergence observed in real networks is robust to different choices for the specific model of human expectations.

7.8.8.3 *Directed networks*

We now consider directed versions of the real networks. Among the 40 networks chosen for analysis, 28 have directed versions (see Tab. 7.13). Analysis of directed networks follows in much the same way as our previous analysis of undirected networks; the only difference is that, when computing the entropy (Eq. 9.1) and KL divergence (Eq. 7.5), we calculate the stationary distribution π numerically by solving the eigenvector equation $\pi^\intercal = \pi^\intercal P$. We find that most of the directed real networks have higher entropy than completely randomized versions (Fig. 7.14a); the main exceptions are the citation networks, which we discuss in further detail in Sec. 7.8.10. We also find that all of the directed real networks have lower KL divergence than completely randomized versions (Fig. 7.14b), where the expectations $\hat{P}$ are calculated using the model in Eq. (7.10)

If we instead compare against randomized versions that preserve both the in- and out-degrees of nodes, we see that the entropy of real networks remains relatively unchanged (Fig. 7.14c); again, the citation networks as a group represent the strongest exception to this result. Even when compared with degree-preserving randomized versions, all of the directed real networks attain a lower KL divergence (Fig. 7.14d). Generally, these results demonstrate that our conclusions regarding the information properties of real networks also apply to directed networks: (i) their entropy is higher than completely

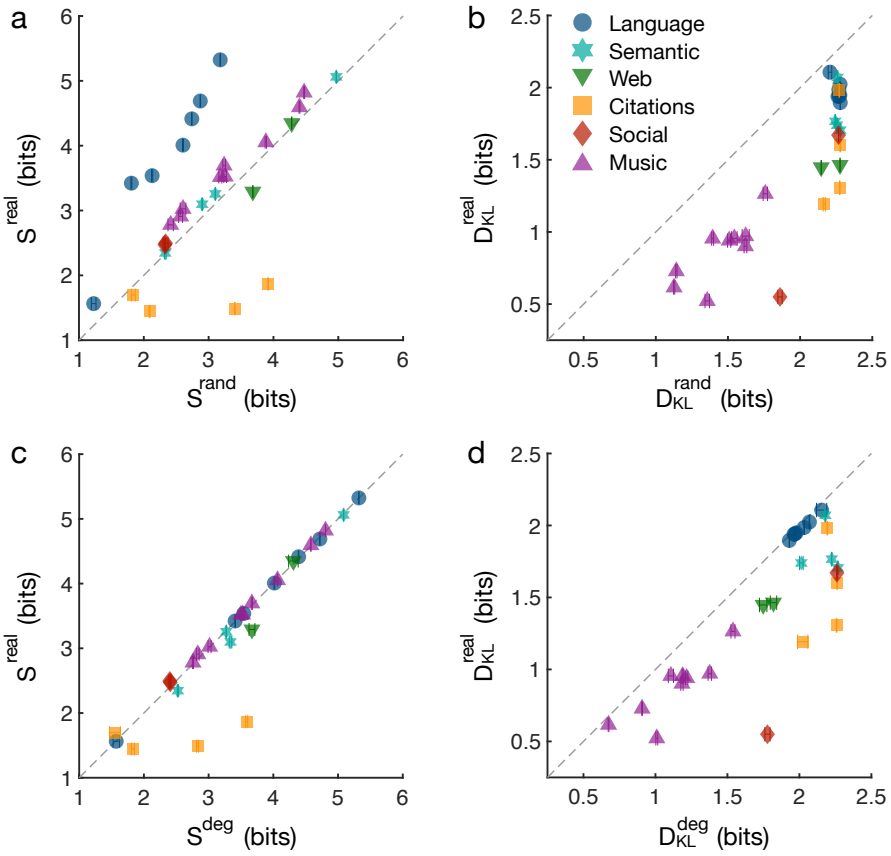


Figure 7.14: Entropy and KL divergence of directed versions of real networks. (*a*) Entropy of directed versions of the real networks listed in Tab. 7.13 (S^{real}) compared with fully randomized versions (S^{rand}). Entropy is calculated directly from Eq. (9.1) with the stationary distribution $\boldsymbol{\pi}$ calculated numerically. (*b*) KL divergence of directed versions of the real networks ($D_{\text{KL}}^{\text{real}}$) compared with fully randomized versions ($D_{\text{KL}}^{\text{rand}}$). Expectations $\hat{P}$ are defined as in Eq. (7.10) with η set to the average value 0.80 from our human experiments. (*c*) Entropy of randomized versions of directed real networks with in- and out-degrees preserved (S^{deg}) compared with S^{real}. (*d*) KL divergence of degree-preserving randomized versions of directed real networks ($D_{\text{KL}}^{\text{deg}}$) compared with $D_{\text{KL}}^{\text{real}}$. Data points and error bars (standard deviations) are estimated from 100 realizations of the randomized networks.

randomized versions and is primarily driven by the degree distribution, and (ii) their KL divergence is lower than both completely randomized and degree-preserving randomized versions.

7.8.9 *Temporally evolving networks*

In the main text, we studied the information properties of static communication networks. However, many of these networks are inherently temporal in nature, evolving over time to arrive at the final form that we observe today (*317*). This observation raises

a number of interesting questions: How does the temporal nature of communication networks affect their ability to communicate information? Moreover, do communication networks evolve over time to optimize efficient communication?

To answer these questions, we consider temporally evolving versions of the real networks studied in the main text. Among the 40 networks chosen for analysis, 19 have temporal versions (see Tab. 7.13), including all of the language (noun transition) and music (note transition) networks, as well as the Facebook network and the two arXiv citation networks. For each network, we record a sequence of up to 100 subnetworks representing different snapshots in the network's evolution. For example, in the language and music networks, each subnetwork represents the transitions between nouns or notes up to a given point in the text or musical piece. Similarly, each subnetwork for the Facebook and citation networks defines the social relationships or scientific citations at a given point in the growth of the corresponding network.

We find that the communication networks maintain higher entropy (Fig. 7.15a) and lower KL divergence (Fig. 7.15b) than completely randomized versions along almost the entirety of their evolutionary processes. Additionally, when compared against degree-preserving randomized versions, we find that the temporally evolving networks have the same entropy (Fig. 7.15c), as expected, and still maintain lower KL divergence along nearly the entire growth process (Fig. 7.15d). These results indicate that, even from the earliest stages in their development, real communication networks are organized to communicate large amounts of information (having high entropy) and to do so efficiently (having low KL divergence from human expectations).

Yet, it remains unclear whether networks evolve over time to optimize efficient communication. To answer this question, we first investigate how the difference between the entropy of real networks and that of fully randomized versions changes over the course of a network's evolution (Fig. 7.16a). Interestingly, across all of the networks considered, we find that this difference in information production increases nearly monotonically as the networks grow, indicating that real communication networks evolve over time to transmit larger and larger amounts of information. Second, we study how the difference between the KL divergence of real networks and that of completely randomized versions changes over the evolution of a network (Fig. 7.16b). Notably, the music, social, and citation networks all evolve over time to minimize this difference, thereby becoming more efficient. However, language networks display a markedly different trajectory, minimizing their KL divergence (relative to randomized versions) until about 10% of the way into their development, and then slowly growing to become less efficient. This pattern indicates that transitions between nouns communicate information most efficiently at the beginning of a text, and then become less efficient (while communicating larger amounts of information) as the text progresses. Together, these results suggest that communication networks evolve to (i) maximize the amount of information being communicated and (ii), with the exception of language networks, minimize the inefficiency of their communication.

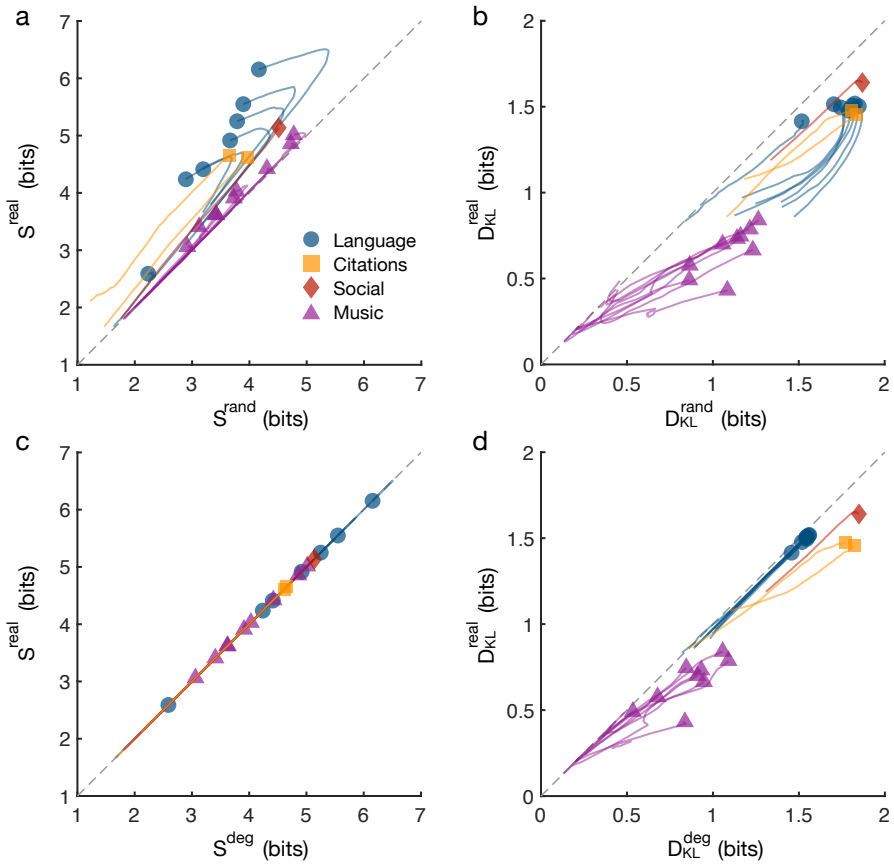


Figure 7.15: Entropy and KL divergence of temporally evolving versions of real networks. (*a*) Entropy of temporally evolving versions of the real networks listed in Tab. 7.13 (S^{real}) compared with fully randomized versions (S^{rand}). Each line represents a sequence of growing networks and each symbol represents the final version of the network. (*b*) KL divergence of evolving versions of the real networks (D_{KL}^{real}) compared with fully randomized versions (D_{KL}^{rand}). Expectations $\hat{P}$ are defined as in Eq. (7.10) with η set to the average value 0.80 from our human experiments. (*c*) Entropy of temporally evolving versions of real networks (S^{real}) compared with degree-preserving randomized versions (S^{deg}). (*d*) KL divergence of temporally evolving versions of real networks (D_{KL}^{real}) compared with degree-preserving randomized versions (D_{KL}^{deg}). Across all panels, each point along the lines represents an average over five realizations of the randomized networks.

7.8.10 *Real networks that do not support efficient communication*

One of the central results of the paper is that real communication networks tend to have two properties: (i) high entropy and (ii) low KL divergence from human expectations. Specifically, these results tend to hold relative to fully randomized and degree-preserving randomized versions of the networks. However, it is useful to consider instances when these general results break down; that is, examples of real communication networks that either have low entropy or high KL divergence. Such examples are important for two reasons: First, they illustrate that efficient communica-

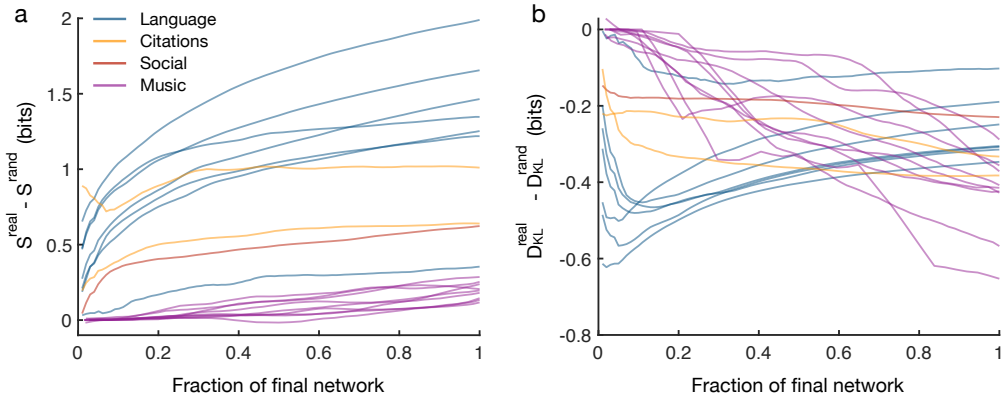


Figure 7.16: Evolution of the difference in entropy and KL divergence between real networks and randomized versions. (*a*) Difference between the entropy of temporally evolving real networks (S^{real}) and the entropy of fully randomized versions of the same networks (S^{rand}) plotted as a function of the fraction of the final network size. Each line represents a sequence of growing networks that culminates in one of the communication networks studied in the main text. (*b*) Difference between the KL divergence of temporally evolving real networks (D_{KL}^{real}) and that of fully randomized versions (D_{KL}^{rand}) plotted as a function of the fraction of the final network size. When calculating the KL divergences, the expectations $\hat{P}$ are defined as in Eq. (7.10) with η set to the average value 0.80 from our human experiments. Across both panels, each point along the lines represents an average over five realizations of the randomized networks.

tion (defined by high entropy and low KL divergence) is not a necessary property of all real-world communication networks; and second, studying their properties reveals how efficient communication can break down. In what follows, we present two examples of real communication networks that do not support the efficient communication of information, either by having low entropy (low information production) or high KL divergence from human expectations (high inefficiency).

7.8.10.1 *Directed citation networks*

First, we consider directed versions of the citation networks studied in the main paper. In the Sec. 7.8.8.3, we found that the directed versions of citation networks have lower entropy than both fully randomized and degree-preserving randomized versions (Fig. 7.14a,c), contradicting our general observation that real communication networks have high entropy. Here we show that this contradiction stems from the inherently temporal nature of citation networks; namely, the fact that directed edges tend to flow backwards in time as more recent papers cite older papers. This temporal feature causes newer papers to have a lower in-degree than older papers, thereby disrupting the natural correlation between in- and out-degree in other real networks. For example, we see in the arXiv Hep-Th citation network that the in- and out-degrees are only weakly correlated (Fig. 7.17a), while for the Shakespeare language network, the in- and out-degrees are tightly correlated (Fig. 7.17b).

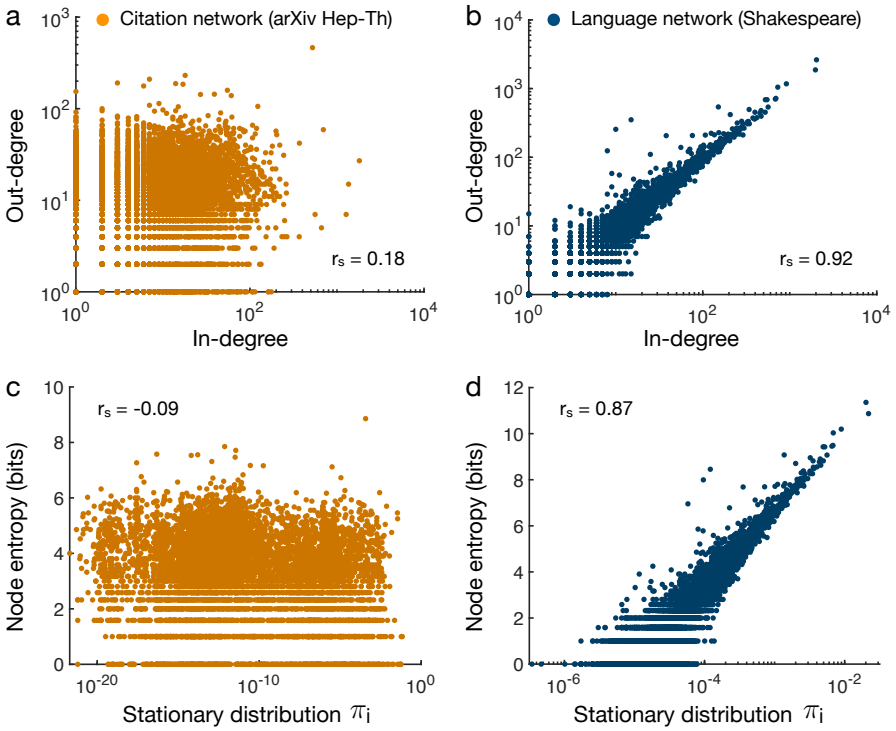


Figure 7.17: Comparison of directed citation and language networks. (*a*) Out-degrees $k_i^{\text{out}} = \sum_j G_{ij}$ of nodes in the arXiv Hep-Th citation network compared with the in-degrees $k_i^{\text{in}} = \sum_j G_{ji}$ of the same nodes; we find a weak Spearman's correlation of $r_s = 0.18$. (*b*) Out-degrees compared with in-degrees of nodes in the Shakespeare language (noun transition) network; we find a strong correlation $r_s = 0.92$. (*c*) Entries in the stationary distribution π_i for different nodes in the citation network compared with the node-level entropy S_i; we find a weakly negative correlation $r_s = -0.09$. (*d*) Entries in the stationary distribution compared with node-level entropies in the language network; we find a strong correlation $r_s = 0.87$.

Since the in-degree of a node i roughly corresponds to the frequency with which random walks visit i, we can think of the in-degrees $\mathbf{k}^{\text{in}}$ as approximately determining the stationary distribution $\boldsymbol{\pi}$. By contrast, the node-level entropy $S_i = -\sum_j P_{ij} \log P_{ij}$ is determined by the out-degree of node i, since $P_{ij} = \frac{1}{k_i^{\text{out}}} G_{ij}$ from Eq. (7.6). Since the network-averaged entropy is simply an inner product of the stationary distribution and the node-level entropy, $S = \sum_i \pi_i S_i$, this quantity is maximized in networks for which π_i and S_i are correlated. Returning to our previous examples, we find that the stationary distribution and node-level entropy are weakly negatively correlated in the citation network (Fig. 7.17c), whereas in the language network, the stationary distribution and node entropy are tightly correlated (Fig. 7.17d). Thus, the apparent contradiction between directed citation networks and our general result that real networks have high entropy is primarily driven by the temporal nature of directed edges in citation networks. Indeed, if one instead allows random walks to flow along either direction of each edge, as in the undirected versions studied in the main text, we find that citation networks do have high entropy (Fig. 7.2a). Therefore, the capacity of

citation networks to communicate large amounts of information depends critically on the ability of walks to hop both forward and backward along citations.

7.8.10.2 *Language networks including all parts of speech*

In the main text, we focus on language networks consisting of the transitions between nouns in a given text. This choice to focus on nouns follows from existing literature that distinguishes "content" words (such as nouns), which contain meaning, from "grammatical" words (such as articles, conjunctions, and prepositions), which define the structure of a sentence (*225*, *447*). If we instead consider language (word transition) networks that include all parts of speech, we find that these all-word transition networks have both higher entropy and lower KL divergence than fully randomized versions, aligning with the results from the main text (Fig. 7.3a,b). However, when compared to degree-preserving randomized versions, we find that the all-word transition networks have nearly the same KL divergence (Fig. 7.18a), with three of the seven networks exhibiting KL divergences that are either higher or statistically indistinguishable from the degree-preserving randomized versions. By contrast, the networks of transitions between nouns studied in the main text all exhibit lower KL divergence than degree-preserving randomized versions (Fig. 7.18b).

From the analytic and numerical results presented in the main text (Fig. 8.2e-h) and Sec. 7.8.12, we know that decreases in KL divergence are largely driven by increases in clustering. Indeed, for the all-word transition networks, we find that the average clustering coefficients are consistently lower than for degree-preserving randomized versions (Fig. 7.18c), thereby explaining their relatively high KL divergences (Fig. 7.18a). To understand the low clustering (and therefore the high KL divergence) of the all-word transition networks, it is helpful to consider the fact that words typically transition from content words to grammatical words in order to maintain grammatical structure. This hopping between content and grammatical words yields transition networks with disassortative community structure (*225*, *447*), wherein words from the same class are less likely to form edges than words in different classes, which, in turn, decreases the clustering. By contrast, if we restrict our attention to content words (such as the nouns studied in the main text), we find that the transition networks exhibit high clustering (Fig. 7.18d) and therefore low KL divergence (Fig. 7.18b).

7.8.11 *Entropy of random walks*

Given the high entropy and low KL divergence from human expectations observed in real networks, it is natural to wonder what topological features give rise to these properties. We note that there has been a large amount of recent research studying maximum entropy random walks, wherein the topology of the network is fixed but the edge weights are tuned to maximize the entropy rate (*117*, *148*, *183*, *615*). By contrast, here we are interested in understanding how, for fixed edge weights, different network topologies either increase or decrease the entropy of random walks.

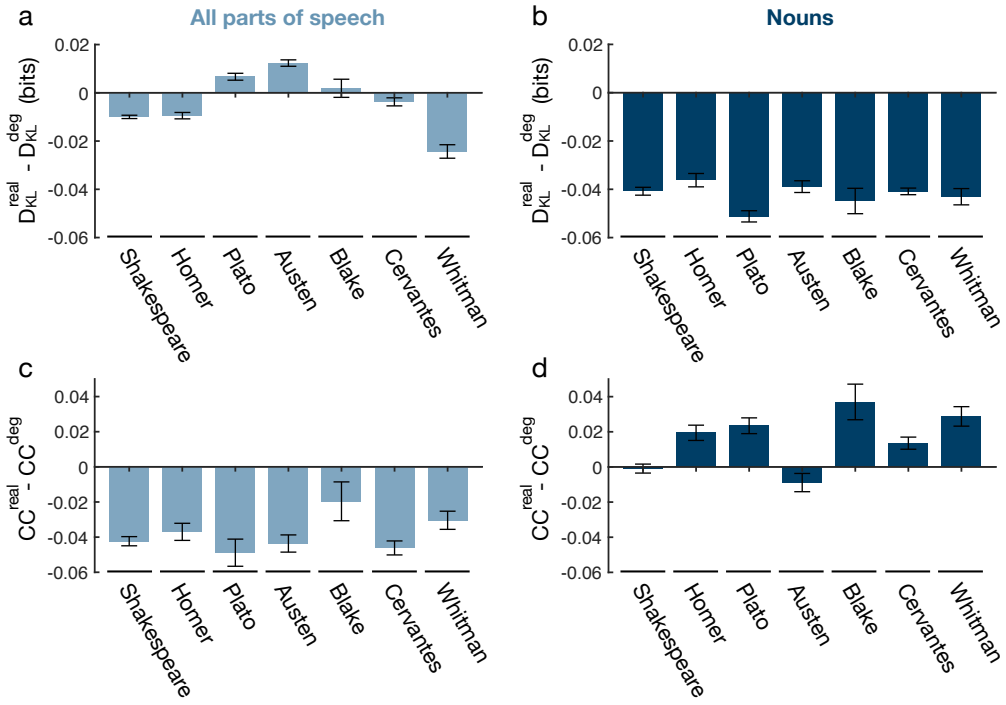


Figure 7.18: Comparison of all-word transition networks and noun transition networks. (*a-b*) Difference between the KL divergence of language (word transition) networks (D_{KL}^{real}) and degree-preserving randomized versions of the same networks (D_{KL}^{deg}). We consider networks of transitions between all words (*a*) and networks of transitions between nouns (*b*). (*c*, *d*) Difference between the average clustering coefficient of language networks (CC^{real}) and degree-preserving randomized versions of the same networks (CC^{deg}), where transitions are considered between all words (*c*) or only nouns (*d*). In all panels, data points and error bars (standard deviations) are estimated from 100 realizations of the randomized networks, and the networks are undirected.

To make analytic progress, we focus on unweighted, undirected networks. In this case, Eq. (7.8) shows that the entropy is determined by the degree sequence of the network. If we consider a random network ensemble with node degrees independently distributed according to a degree distribution $\mathcal{P}(k)$, then the average entropy rate is given by (*263*)

$$\begin{aligned} \langle S \rangle &= \frac{1}{2E} \sum_i \langle k_i \log k_i \rangle \\ &= \frac{\langle k \log k \rangle}{\langle k \rangle}, \end{aligned} \tag{7.12}$$

where the averages are taken over $\mathcal{P}(k)$.

7.8.11.1 *High-degree expansion*

Since $k \log k$ is convex in k, it is clear that $\langle k \log k \rangle \geqslant \langle k \rangle \log \langle k \rangle$, and we arrive at a simple lower bound for the entropy,

$$\langle S \rangle \geqslant \log \langle k \rangle. \tag{7.13}$$

In fact, one can show that $\log \langle k \rangle$ is the zeroth-order term in an expansion of $\langle S \rangle$ in the limit of large average degree $\langle k \rangle \gg 1$. Expanding $k \log k$ around $\langle k \rangle$, we find

$$\begin{aligned} \langle S \rangle &= \frac{1}{\langle k \rangle} \langle \langle k \rangle \log \langle k \rangle + (1 + \log \langle k \rangle)(k - \langle k \rangle) + \frac{(k - \langle k \rangle)^2}{2\langle k \rangle} + O\left(\frac{1}{\langle k \rangle^2}\right) \rangle \\ &= \log \langle k \rangle + \frac{\mathrm{Var}(k)}{2\langle k \rangle^2} + O\left(\frac{1}{\langle k \rangle^3}\right), \end{aligned} \tag{7.14}$$

where $\mathrm{Var}(k)$ is the variance of k. We therefore find that, in addition to increasing logarithmically with the average degree, the entropy of random walks grows with increasing degree variance. In turn, this result further supports the conclusion that networks with heterogeneous degrees produce random walks with higher entropy. In what follows, we derive analytic results for the entropy of random walks on various canonical network families.

7.8.11.2 k-*regular network*

We begin by studying k-regular networks, wherein each node i has constant degree $k_i = k$. In this case, we arrive at the simple relation $\langle S \rangle = \log k$, which saturates the lower bound in Eq. (7.13) (*161*). This result shows that k-regular networks achieve the lowest possible entropy among networks of a given density.

7.8.11.3 *Poisson distributed network*

While k-regular networks maintain a lattice-like structure, many real networks display random organization (*11*). The simplest model for generating random networks, known as the Erdös-Rényi model (*204*), places E edges uniformly at random between pairs of N nodes. In the thermodynamic limit $N \to \infty$, Erdös-Rényi networks follow a Poisson degree distribution $\mathcal{P}(k) = e^{-\langle k \rangle} \langle k \rangle^k / k!$. In this case, the degree variance is given by $\mathrm{Var}(k) = \langle k \rangle$, and applying Eq. (7.14), we find that

$$\langle S \rangle = \log \langle k \rangle + \frac{1}{2\langle k \rangle} + O\left(\frac{1}{\langle k \rangle^2}\right). \tag{7.15}$$

Therefore, in the high-$\langle k \rangle$ limit, the entropy of random walks on an Erdös-Rényi network approaches the lower-bound $\langle S \rangle \approx \log \langle k \rangle$. We find that the analytic prediction in Eq. (7.15) accurately approximates the true entropy of randomly-generated Erdös-Rényi networks across all values of the average degree (Fig. 7.19a).

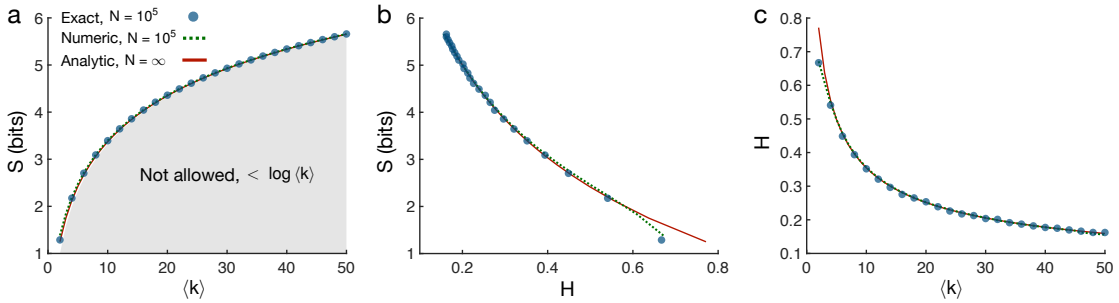


Figure 7.19: Entropy of random walks in Poisson distributed networks. (*a*) Entropy of random walks as a function of the average degree $\langle k \rangle$ for Poisson distributed networks. Data points are exact calculations using the degree sequences of randomly-generated Erdös-Rényi networks of size $N = 10^4$. Dashed lines are numerical results for $N = 10^4$, calculated using the Poisson degree distribution. Solid lines are analytic results for the thermodynamic limit $N \to \infty$. (*b*) Entropy as a function of the degree heterogeneity H for variable $\langle k \rangle$. (*c*) Degree heterogeneity as a function of the average degree.

To investigate the relationship between the entropy and the heterogeneity of degrees in a network, we defined the degree heterogeneity to be the relative average difference in degrees,

$$H = \frac{\langle |k_i - k_j| \rangle}{\langle k \rangle} = \frac{1}{\langle k \rangle} \sum_{k_i, k_j} |k_i - k_j| \mathcal{P}(k_i) \mathcal{P}(k_j). \tag{7.16}$$

H is a well-studied measure of the dispersion of a distribution, with range $[0, 2]$. We note that other often used measures of degree heterogeneity, such as $\langle k^2 \rangle / \langle k \rangle^2$ and $\mathrm{Var}(k)/\langle k \rangle^2$, cannot be used to study the impact of degree heterogeneity on entropy for scale-free networks since $\langle k^2 \rangle$ diverges for $\gamma \leqslant 3$ in the limit $N \to \infty$. For Poisson distributed networks, one can show that

$$H = 2e^{-2\langle k \rangle} \big(I_0(2\langle k \rangle) + I_1(2\langle k \rangle) \big), \tag{7.17}$$

where $I_\nu(x)$ is the modified Bessel function of the first kind (*410*). For other degree distributions, however, it is generally difficult to derive an analytic form for H. We find that the entropy of random walks on Poisson distributed networks decreases with increasing degree heterogeneity as we vary $\langle k \rangle$ (Fig. 7.19b), seemingly contradicting our conclusion in the main text that entropy increases with heterogeneity. However, this effect is driven by the monotonic decrease in H with increasing $\langle k \rangle$ in Poisson distributed networks (Fig. 7.19c). In the following subsections, we show that entropy does in fact increase with degree heterogeneity for other network models, confirming the results in the main text.

7.8.11.4 *Power-law distributed network*

Compared to random networks, real networks often contain a number of hub nodes with unusually high degree, leading to a heavy-tailed distribution of node degrees

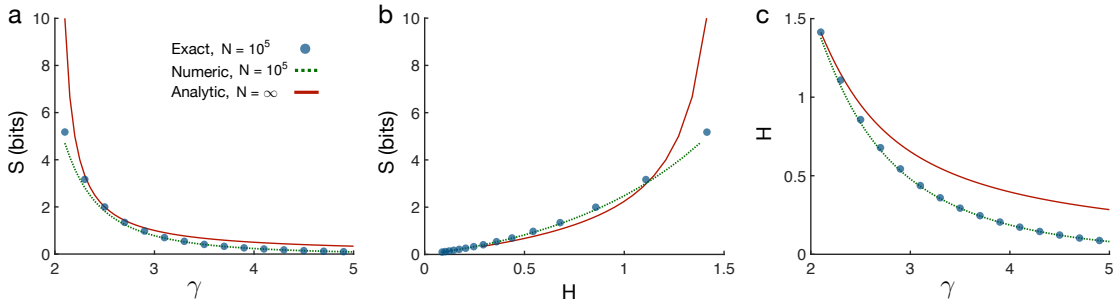


Figure 7.20: Entropy of random walks in power-law distributed networks. (*a*) Entropy of random walks as a function of the scale-free exponent γ for power-law distributed networks. Data points are exact calculations from networks of size $N = 10^4$ generated using the configuration model (*450*). Dashed lines are numerical results for $N = 10^4$, calculated using the power-law degree distribution. Solid lines are analytic results for the thermodynamic limit $N \to \infty$. (*b*) Entropy as a function of the degree heterogeneity H for variable γ. (*c*) Degree heterogeneity as a function of the scale-free exponent.

(*11*). Often this heavy-tailed distribution is associated with scale-free organization (*50*), which is characterized by a power-law degree distribution $P(k) \sim k^{-\gamma}$, where γ is the scale-free exponent. In the limit $N \to \infty$, we can approximate the averages in Eq. (7.12) as integrals, and one can show that (*263*)

$$\langle S \rangle = \frac{1}{\gamma - 2}. \tag{7.18}$$

We see that the entropy diverges as $\gamma \to 2$, while for $\gamma > 2$ the entropy of scale-free networks is well-defined. We remark that this critical exponent is different from $\gamma = 3$, which is the critical exponent for many other network phenomena driven by the divergence of $\langle k^2 \rangle$ (*13*, *513*). Instead, as $\gamma \to 2$, super-hubs emerge that connect to almost all of the nodes in the network, causing the average degree $\langle k \rangle$ to diverge (*11*). Each time a random walk arrives at one of these super-hubs, the entropy of the ensuing transition, roughly $-\log \frac{1}{N}$, diverges as $N \to \infty$.

We compare the analytic prediction in Eq. (7.18) with exact calculations from both power-law distributed networks generated using the configuration model (*450*) and from numerical calculations of the averages in Eq. (7.12), finding that the numerical estimates agree well with the exact values (Fig. 7.20a). Moreover, we find that the entropy increases with degree heterogeneity as we sweep over γ (Fig. 7.20b), confirming our conclusions in the main text. This increase in entropy is related to the corresponding increase in heterogeneity as $\gamma \to 2$ (Fig. 7.20c).

7.8.11.5 *Static model*

In order to test the effects of network density and degree heterogeneity independently, we turn to the static model, which is commonly used to generate scale-free networks of a given density (*258*). Beginning with N disconnected nodes, we assign each node

i a weight $w_i = i^{-\alpha}$, where $\alpha \in [0, 1)$ is a real number. Then, we randomly select a pair of nodes i and j with probabilities proportional to their weights, and we connect them if they have not already been connected. This process is repeated until $E = \frac{1}{2}N\langle k\rangle$ edges have been added. A number of analytic properties have been derived for the static model (*128*, *392*), including the fact that, in the thermodynamic limit, the degree distribution is given by $\mathcal{P}(k) = \frac{1}{\alpha}\left(\frac{\langle k\rangle}{2}(1-\alpha)\right)^{1/\alpha} \frac{\Gamma\left(k-\frac{1}{\alpha}, \frac{\langle k\rangle}{2}(1-\alpha)\right)}{\Gamma(k+1)}$, where $\Gamma(\cdot)$ is the gamma function and $\Gamma(\cdot, \cdot)$ is the upper incomplete gamma function. In the large-k limit, one can show that the degree distribution drops off as a power law $\mathcal{P}(k) \sim k^{-\gamma}$, where $\gamma = 1 + \frac{1}{\alpha}$.

We are interested in deriving an analytic form for the entropy. Using a hidden variables method (*128*), one can show that the average degree of node i is given by

$$\bar{k}(i) = \langle k\rangle(1-\alpha)\left(\frac{i}{N}\right)^{-\alpha}\left(1 - N^{\alpha-1}\right). \tag{7.19}$$

Approximating the numerator in Eq. (7.12) by $\langle k \log k\rangle \approx \frac{1}{N}\int_1^N \bar{k}(i)\log\bar{k}(i)\,di$, and taking the limit $N \to \infty$, we find that the entropy is given by

$$\langle S\rangle = \log\langle k\rangle + \frac{1}{\gamma-2} - \log\frac{\gamma-1}{\gamma-2}. \tag{7.20}$$

We note that the average degree of a pure power-law network is $\langle k\rangle = \frac{\gamma-1}{\gamma-2}$. Plugging this average degree into Eq. (7.20), we recover the entropy of power-law distributed networks in Eq. (7.18), as expected. Interestingly, we notice that, even for finite $\langle k\rangle$, the entropy in the static model diverges as $\gamma \to 2$ in the thermodynamic limit.

We find that the entropy increases as $\langle k\rangle$ increases (Fig. 7.21a) and also as γ decreases (Fig. 7.21b). The thermodynamic result in Eq. (7.20) is accurate for $\gamma \geqslant 3$, while numerical calculations using Eq. (7.19) and including finite network size yield accurate predictions for $\gamma \geqslant 2.5$. We note that the only effect of $\langle k\rangle$ on the entropy in Eq. (7.20) is in the logarithmic lower bound, suggesting that the quantity $S - \log\langle k\rangle$ should depend exclusively on the scale-free exponent γ. Indeed, subtracting $\log\langle k\rangle$ from our entropy calculations, we find that networks of varying density collapse onto a single line (Fig. 7.21c). This result is made even more clear by considering how the quantity $S - \log\langle k\rangle$ varies with degree heterogeneity as we sweep over γ (Fig. 7.21e). Finally, we note that H increases with decreasing γ (Fig. 7.21f), thereby explaining the monotonic relationship between entropy and degree heterogeneity in the static model (Fig. 7.21d).

7.8.11.6 *Exponentially distributed network*

Many real networks exhibit degree distributions with exponential cutoffs for large values of k (*11*, *477*). In pure exponentially distributed networks, the degree distribution follows the form $\mathcal{P}(k) \sim e^{-k/\kappa}$, where $\kappa \geqslant 0$ is the degree cutoff. In the thermodynamic

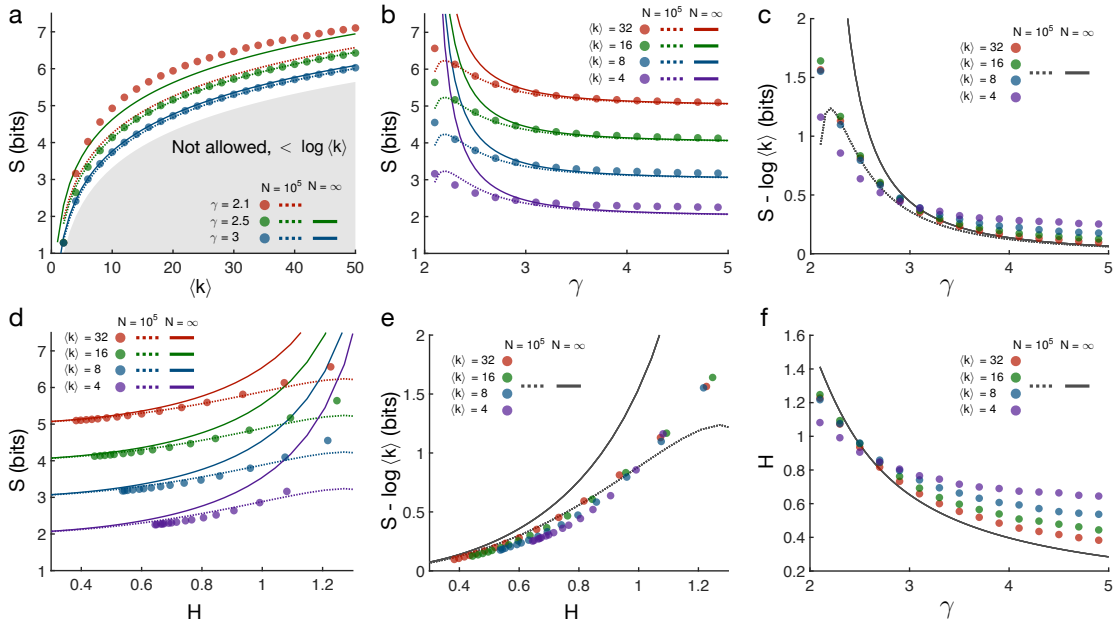


Figure 7.21: Entropy of random walks in static model networks. (*a*) Entropy of random walks as a function of the average degree $\langle k \rangle$ for various values of the scale-free exponent γ in the static model. Data points are exact calculations using the degree sequences of networks with $N = 10^4$ generated using the static model. Dashed lines are numerical results for $N = 10^4$, calculated using the average degree relationship in Eq. (7.19). Solid lines are analytic results for the thermodynamic limit $N \to \infty$. (*b*) Entropy as a function of γ for various values of $\langle k \rangle$. (*c*) The quantity $S - \log\langle k \rangle$ collapses to a single function of γ across various values of $\langle k \rangle$. (*d*) Entropy as a function of the degree heterogeneity H for varying γ. (*e*) The quantity $S - \log\langle k \rangle$ increases with H for varying γ. (*f*) Degree heterogeneity increases as γ decreases toward the critical value $\gamma = 2$.

limit, approximating the averages in Eq. (7.12) as integrals, we find that the entropy is given by

$$\langle S \rangle = \log\langle k \rangle + \frac{1 - \gamma_e}{\ln 2}, \tag{7.21}$$

where γ_e is Euler's constant. We see in Fig. 7.22a that this analytic prediction accurately describes the entropy of randomly-generated exponential networks. Moreover, we find that the entropy increases with increasing degree heterogeneity (Fig. 7.22b) and that the heterogeneity increases with the degree cutoff κ (Fig. 7.22c).

7.8.12 *KL divergence between random walks and human expectations*

The results of the previous section demonstrate that, generally, the entropy of random walks is larger for networks with heterogeneous degrees, a feature that has been found in many real networks (*50*, *51*, *121*, *477*). But what are the structural features that allow a network to maintain a low divergence from human expectations? Here, we answer

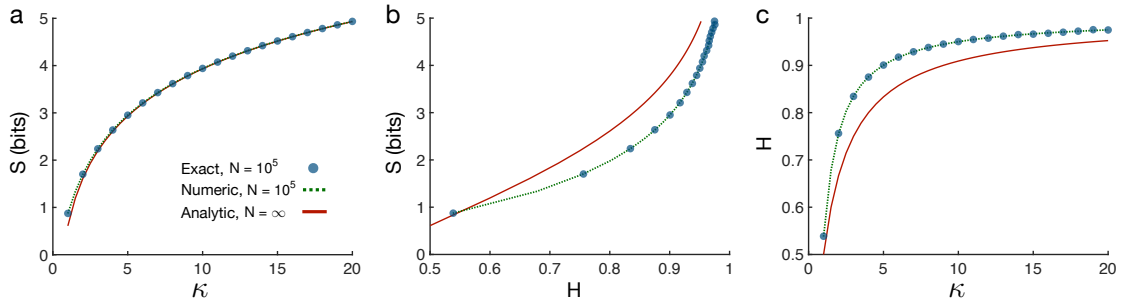


Figure 7.22: Entropy of random walks in exponentially distributed networks. (*a*) Entropy of random walks as a function of the degree cutoff κ for exponentially distributed networks. Data points are exact calculations from networks of size $N = 10^4$ generated using the configuration model (*450*). Dashed lines are numerical results for $N = 10^4$, calculated using the exponential degree distribution. Solid lines are analytic results for the thermodynamic limit $N \to \infty$. (*b*) Entropy as a function of the degree heterogeneity for variable κ. (*c*) Degree heterogeneity as a function of the exponential cutoff.

this question by studying the KL divergence $D_{KL}(P||\hat{P})$ between a network's transition structure P and the expectations of an observer $\hat{P}$.

7.8.12.1 *Upper bound*

For expectations $\hat{P}$ of the form in Eq. (7.9), the KL divergence is given by

$$\begin{aligned} D_{KL}(P||\hat{P}) &= -\sum_i \pi_i \sum_j P_{ij} \log \frac{\hat{P}_{ij}}{P_{ij}} \\ &= -\sum_i \pi_i \sum_j P_{ij} \log \left(C \sum_{t=0}^{\infty} f(t) \frac{(P^{t+1})_{ij}}{P_{ij}} \right), \end{aligned} \tag{7.22}$$

where $(P^{t+1})_{ij}/P_{ij}$ is the relative probability of transitioning from node i to node j in $t+1$ steps versus one step. Keeping only the first term inside the logarithm, we arrive at an upper bound for the KL divergence,

$$D_{KL}(P||\hat{P}) \leqslant -\sum_i \pi_i \sum_j P_{ij} \log\left(Cf(0)\right) = -\log\left(Cf(0)\right). \tag{7.23}$$

Eq. (7.23) allows us to make a number of simple predictions for the KL divergence. For example, if the expectations are defined by $f(t) = \eta^t$, as presented in the main text, then $C = 1-\eta$ and so $D_{KL} \leqslant -\log(1-\eta)$. In this case, we see that the KL divergence tends to zero as $\eta \to 0$ and that the upper bound diverges as $\eta \to 1$. In contrast, if $f(t) = (t+1)^{-\alpha}$ then $C = \zeta(\alpha)^{-1}$, where $\zeta(\cdot)$ is the Riemann zeta function, and we have $D_{KL} \leqslant \log \zeta(\alpha)$. As a final example, if $f(t) = 1/t!$ then $C = e^{-1}$, and so $D_{KL} \leqslant \log e$ (which we remark is not equal to one since we use log base two).

7.8.12.2 *Relationship to clustering*

While Eq. (7.23) provides a simple relationship between the KL divergence and parameters in the model for $\hat{P}$, we are ultimately interested in understanding the effects of network structure. To gain an intuition for the role of topology, it helps to focus on a particular model for the expectations. For example, considering $f(t) = \eta^t$, in the low-η limit the KL divergence takes the form

$$\begin{aligned} D_{KL}(P||\hat{P}) &= -\log(1-\eta) - \sum_i \pi_i \sum_j P_{ij} \log\left(1 + \eta \frac{(P^2)_{ij}}{P_{ij}} + O(\eta^2)\right) \\ &= -\log(1-\eta) - \frac{\eta}{\ln 2} \sum_i \pi_i \sum_j P_{ij} \frac{(P^2)_{ij}}{P_{ij}} + O(\eta^2). \end{aligned} \tag{7.24}$$

We note that, when calculating information measures such as entropy or KL divergence, one only considers terms with non-zero probability, such that, for each node i, the sum on j in Eq. (7.24) implicitly runs over all nodes for which $P_{ij} = \frac{1}{k_i} G_{ij}$ is non-zero. Therefore, for undirected networks, recalling that $\pi_i = \frac{k_i}{2E}$, we have

$$D_{KL}(P||\hat{P}) = -\log(1-\eta) - \frac{\eta}{2E \ln 2} \sum_i k_i \sum_j G_{ij} \sum_\ell \left(\frac{1}{k_i} G_{i\ell}\right)\left(\frac{1}{k_\ell} G_{\ell j}\right) + O(\eta^2). \tag{7.25}$$

Switching the i and ℓ indices and canceling terms, we arrive at the concise approximation

$$D_{KL}(P||\hat{P}) = -\log(1-\eta) - \frac{\eta}{E \ln 2} \sum_i \frac{1}{k_i} \triangle_i + O(\eta^2), \tag{7.26}$$

where $\triangle_i = (G^3)_{ii}/2$ is the number of (possibly weighted) triangles involving node i. We therefore find that the KL divergence is lower for networks with a lager number of triangles or, equivalently, a higher clustering coefficient. In the following subsections, we investigate the relationship between KL divergence and clustering in Erdös-Rényi and stochastic block networks.

7.8.12.3 *Erdös-Rényi network*

We wish to derive an analytic approximation for the KL divergence of an Erdös-Rényi network. Considering human expectations defined by $f(t) = \eta^t$, for undirected networks Eq. (7.22) becomes

$$\begin{aligned} D_{KL}(P||\hat{P}) &= -\sum_i \frac{k_i}{2E} \sum_j \frac{1}{k_i} G_{ij} \log\left((1-\eta)\sum_{t=0}^{\infty} \eta^t \frac{(P^{t+1})_{ij}}{P_{ij}}\right) \\ &= -\log(1-\eta) - \frac{1}{2E} \sum_{ij} G_{ij} \log\left(\sum_{t=0}^{\infty} \eta^t \frac{(P^{t+1})_{ij}}{P_{ij}}\right). \end{aligned} \tag{7.27}$$

We note that the second term above is an average of the logarithm over the edges in the network. Approximating this average of logarithms by a logarithm of the average, we have

$$D_{KL}(P||\hat{P}) \approx -\log(1-\eta) - \log\left[\frac{1}{2E}\sum_{ij} k_i G_{ij} \sum_{t=0}^{\infty} \eta^t (P^{t+1})_{ij}\right]. \tag{7.28}$$

For unweighted networks, we recognize that $\sum_j G_{ij}(P^{t+1})_{ij}$ is the probability of transitioning from node i to one of i's neighbors in $t+1$ steps. For $t = 0$ this probability is one. For $t > 0$, we consider two cases: (i) dense networks with high $\langle k\rangle$, and (ii) sparse networks with low $\langle k\rangle$.

For dense networks, we approximate the probability of transitioning from node i to one of node i's neighbors in $t+1 > 1$ steps as k_i/N, the probability of randomly selecting one of the k_i neighbors from all N nodes. Plugging this approximation for $\sum_j G_{ij}(P^t)_{ij}$ into Eq. (7.28), we have

$$\begin{aligned} D_{KL}(P||\hat{P}) &\approx -\log(1-\eta) - \log\left[\frac{1}{2E}\sum_i k_i \left(1 + \sum_{t=1}^{\infty} \eta^t \frac{k_i}{N}\right)\right] \\ &= -\log(1-\eta) - \log\left[1 + \frac{1}{2EN}\frac{\eta}{1-\eta}\sum_i k_i^2\right]. \end{aligned} \tag{7.29}$$

We have now reduced the KL divergence to a function of the degree sequence $\mathbf{k}$. For large Erdös-Rényi networks, the node degrees follow a Poisson distribution, and, for large $\langle k\rangle$, we have $\langle k^2\rangle \approx \langle k\rangle^2$. Thus, the average KL divergence for a dense Erdös-Rényi network can be approximated by

$$\begin{aligned} \langle D_{KL}\rangle &\approx -\log(1-\eta) - \langle \log\left[1 + \frac{1}{2E}\frac{\eta}{1-\eta}k^2\right]\rangle \\ &\approx -\log(1-\eta) - \log\left[1 + \frac{1}{2E}\frac{\eta}{1-\eta}\langle k^2\rangle\right] \\ &\approx -\log(1-\eta) - \log\left[1 + \frac{1}{2E}\frac{\eta}{1-\eta}\langle k\rangle^2\right] \\ &= -\log\left[1 - \eta\left(1 - \frac{\langle k\rangle}{N}\right)\right], \end{aligned} \tag{7.30}$$

where the averages are taken over the degree distribution $\mathcal{P}(k)$. We find that this approximation accurately predicts the KL divergence of Erdös-Rényi networks as a function of the integration parameter η (Fig. 7.23a). We also see that in the thermodynamic limit $N \to \infty$, D_{KL} approaches the upper bound $-\log(1-\eta)$.

For sparse Erdös-Rényi networks, the number of loops is small and thus the network is locally treelike (*359*). In a tree, the probability $\sum_j G_{ij}(P^t)_{ij}$ of transitioning from a given node i to one of node i's neighbors in $t+1$ steps is zero if t is odd. For t even,

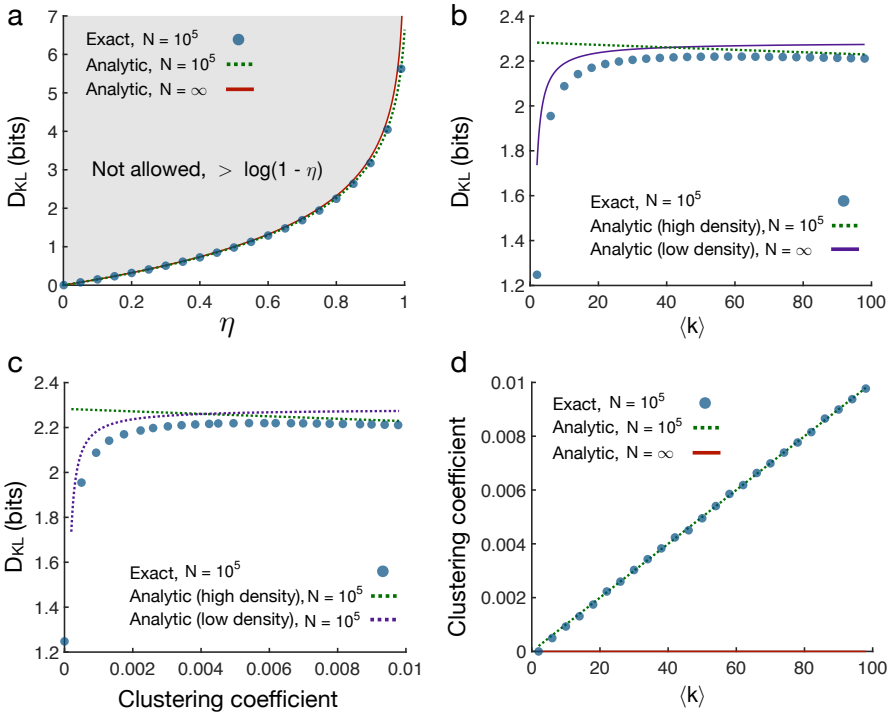


Figure 7.23: KL divergence from human expectations in Erdös-Rényi networks. (*a*) KL divergence between random walks and human expectations as a function of the inaccuracy parameter η for Erdös-Rényi networks. Data points are exact calculations for networks of size $N = 10^4$ with average degree $\langle k \rangle = 100$. Dashed line is the analytic prediction using Eq. (7.30) with $N = 10^4$. Solid line is the analytic result for the thermodynamic limit $N \to \infty$. (*b*) KL divergence as a function of the average degree $\langle k \rangle$ for η equal to the value 0.80 measured in the serial response experiments. Dashed line represents the high-density analytic approximation in Eq. (7.30) with $N = 10^4$, while the solid line is the low-density approximation in Eq. (7.32). (*c*) KL divergence as a function of the average clustering coefficient for variable $\langle k \rangle$. (*d*) Average clustering coefficient as a function of $\langle k \rangle$. In the thermodynamic limit the clustering tends toward zero for all values of $\langle k \rangle$ (solid line).

setting node i to be the root of the tree, if we assume all nodes have the same degree $\langle k \rangle$, then the probability of moving down the tree on any given step is $1 - 1/\langle k \rangle$ and the probability of moving up the tree is $1/\langle k \rangle$. Approximating $1 - 1/\langle k \rangle \approx 1$, then the

probability of moving down the tree $t/2+1$ steps and back up the tree $t/2$ steps is roughly $1/\langle k\rangle^{\frac{t}{2}}$. Plugging this expression into Eq. (7.28), we have

$$\begin{aligned} D_{KL}(P||\hat{P}) &\approx -\log(1-\eta) - \log\left[\frac{1}{2E}\sum_i k_i \sum_{t\ \text{even}} \eta^t \langle k\rangle^{-\frac{t}{2}}\right] \\ &= -\log(1-\eta) - \log\left[\frac{1}{2E}\sum_i k_i \sum_{t=0}^{\infty} \eta^{2t}\langle k\rangle^{-t}\right] \\ &= -\log(1-\eta) - \log\left[\frac{1}{2E}\frac{\langle k\rangle}{\langle k\rangle-\eta^2}\sum_i k_i\right]. \end{aligned} \tag{7.31}$$

Averaging over the Poisson degree distribution, we have

$$\begin{aligned} \langle D_{KL}\rangle &\approx -\log(1-\eta) - \left\langle \log\left[\frac{N}{2E}\frac{\langle k\rangle}{\langle k\rangle-\eta^2}k\right]\right\rangle \\ &\approx -\log(1-\eta) - \log\left[\frac{\langle k\rangle}{\langle k\rangle-\eta^2}\right]. \end{aligned} \tag{7.32}$$

We find that the above approximation provides a decent estimate of the KL divergence for low $\langle k\rangle$, while the high-density approximation in Eq. (7.30) accurately predicts the KL divergence for $\langle k\rangle > 50$ (Fig. 7.23b).

In addition to the dependence of D_{KL} on η and $\langle k\rangle$, we are also interested in the effect of clustering. The clustering coefficient of a given node i is the number of triangles $\triangle_i$ involving node i divided by the number of possible triangles $\binom{k_i}{2} = k_i(k_i-1)/2$. For Erdös-Rényi networks, averaging over all nodes i, the clustering coefficient is approximately $\langle k\rangle/N$. We find that, for small $\langle k\rangle$, the KL divergence increases with increasing clustering, while, for large $\langle k\rangle$, the KL divergence decreases (Fig. 7.23c). Given that the clustering is directly proportional to $\langle k\rangle$ in Erdös-Rényi networks (Fig. 7.23d), the effects of clustering on D_{KL} are driven by the density of the network. To disambiguate the effects of clustering and density, in the following subsection, we study a stochastic block model in which these properties can be varied independently.

7.8.12.4 *Stochastic block network*

In order to test the effects of clustering on the KL divergence without the confounding impact of edge density, we consider the stochastic block model (*177*). Specifically, the N nodes are divided into N/N_c communities of N_c nodes each. Then, a prescribed fraction f of the $E = \langle k\rangle N/2$ edges are placed between pairs of nodes within the same community, and the remaining fraction $1-f$ of edges are placed between nodes in different communities.

We wish to understand the dependence of the KL divergence on the fraction f of within-community edges. Beginning with Eq. (7.28), we once again consider the

probability $\sum_j G_{ij}(P^{t+1})_{ij}$ of transitioning from node i to one of node i's neighbors in $t+1$ steps. As before, for $t = 0$ this probability is one. For $t > 0$, we approximate

$$\sum_j G_{ij}(P^{t+1})_{ij} \approx p^{\text{in}}(t+1)\frac{k_i^{\text{in}}}{N_c - 1} + p^{\text{out}}(t+1)\frac{k_i^{\text{out}}}{N - N_c}, \tag{7.33}$$

where $p^{\text{in}}(t+1)$ is the probability of ending up in the same community as node i after $t+1$ steps, $p^{\text{out}}(t+1)$ is the probability of ending up in a different community from node i after $t+1$ steps, $k_i^{\text{in}} \approx f k_i$ is the number of edges connecting node i to nodes within the same community, and $k_i^{\text{out}} \approx (1-f)k_i$ is the number of edges connecting node i with nodes in different communities. We model the transitions in and out of node i's community as a two-state Markov process with probability matrix

$$A = \begin{pmatrix} P(\text{in} \mid \text{in}) & P(\text{in} \mid \text{out}) \\ P(\text{out} \mid \text{in}) & P(\text{out} \mid \text{out}) \end{pmatrix} = \begin{pmatrix} f & 1-f \\ \frac{N_c}{N-N_c}(1-f) & f + \frac{N-2N_c}{N-N_c}(1-f) \end{pmatrix}. \tag{7.34}$$

Using this representation, one can show that

$$\begin{aligned} p^{\text{in}}(t+1) = (A^{t+1})_{11} &\approx \frac{1}{N}\left(N_c + (N-N_c)f^{t+1}\right), \\ \text{and} \quad p^{\text{out}}(t+1) = (A^{t+1})_{12} &\approx \frac{N-N_c}{N}(1-f^{t+1}), \end{aligned} \tag{7.35}$$

where the approximations follow from the assumption that $\left(\frac{fN-N_c}{N-N_c}\right)^{t+1} \approx f^{t+1}$. Plugging Eq. (7.35) into Eq. (7.33), we have

$$\begin{aligned} \sum_j G_{ij}(P^{t+1})_{ij} &\approx \frac{fk_i}{N}\left(1 + f^{t+1}\left(\frac{N}{N_c} - 1\right)\right) + \frac{(1-f)k_i}{N}\left(1 - f^{t+1}\right) \\ &= \frac{k_i}{N}\left(1 + f^{t+1}\left(\frac{N}{N_c}f - 1\right)\right) \\ &\approx \frac{k_i}{N}\left(1 + \frac{N}{N_c}f^{t+2}\right), \end{aligned} \tag{7.36}$$

where the final approximation follows from the assumption that $\frac{N}{N_c} f \gg 1$. We substitute this result into Eq. (7.28), finding that

$$\begin{aligned} D_{KL}(P||\hat{P}) &\approx -\log(1-\eta) - \log\left[\frac{1}{2E}\sum_i k_i\left(1 + k_i \sum_{t=1}^{\infty} \eta^t \left(\frac{1}{N} + \frac{f^{t+2}}{N_c}\right)\right)\right] \\ &= -\log(1-\eta) - \log\left[1 + \frac{1}{2E}\sum_i k_i^2 \sum_{t=1}^{\infty} \eta^t \left(\frac{1}{N} + \frac{f^{t+2}}{N_c}\right)\right] \\ &= -\log(1-\eta) - \log\left[1 + \frac{1}{2E}\left(\frac{1}{N}\frac{\eta}{1-\eta} + \frac{1}{N_c}\frac{\eta f^3}{1-\eta f}\right)\sum_i k_i^2\right] \\ &= -\log\left[1 - \eta + \frac{\eta}{2E}\left(\frac{1}{N} + \frac{1}{N_c}\frac{(1-\eta)f^3}{1-\eta f}\right)\sum_i k_i^2\right]. \end{aligned} \tag{7.37}$$

For stochastic block models in the thermodynamic limit $N \to \infty$, the degree distribution is Poisson, and for large $\langle k \rangle$ we have $\langle k^2 \rangle \approx \langle k \rangle^2$. Averaging over the Poisson degree distribution, the average KL divergence can be approximated by

$$\begin{aligned} \langle D_{KL} \rangle &\approx -\left\langle \log\left[1 - \eta + \frac{\eta}{2E}\left(\frac{1}{N} + \frac{1}{N_c}\frac{(1-\eta)f^3}{1-\eta f}\right)\sum_i k_i^2\right]\right\rangle \\ &\approx -\log\left[1 - \eta + \frac{\eta N}{2E}\left(\frac{1}{N} + \frac{1}{N_c}\frac{(1-\eta)f^3}{1-\eta f}\right)\langle k^2 \rangle\right] \\ &\approx -\log\left[1 - \eta\left(1 - \frac{\langle k \rangle}{N} - \frac{\langle k \rangle}{N_c}\frac{(1-\eta)f^3}{1-\eta f}\right)\right]. \end{aligned} \tag{7.38}$$

We remark that the first three terms inside the logarithm in Eq. (7.38) are identical to the Erdös-Rényi result in Eq. (7.30), and thus the final term can be regarded as a correction resulting from the modular structure of the stochastic block model. Interestingly, this third term does not vanish in the thermodynamic limit $N \to \infty$; however, it does vanish in the limit $f \to 0$, as the network loses its block structure. We find that the analytic prediction in Eq. (7.38) is accurate across all values of η and all fractions f (Fig. 7.24a,b). Furthermore, we find that the KL divergence decreases monotonically with increasing f for fixed average degree $\langle k \rangle$ (Fig. 7.24a,b).

In order to predict the effect of clustering, it is helpful to have an analytic approximation for the average clustering coefficient in a stochastic block network. We recall that the clustering coefficient for a node i is given by $2\triangle_i / (k_i(k_i - 1))$, where $\triangle_i$ is the number of triangles involving node i. For a stochastic block network, we define the probability of an edge existing between two nodes in the same community as

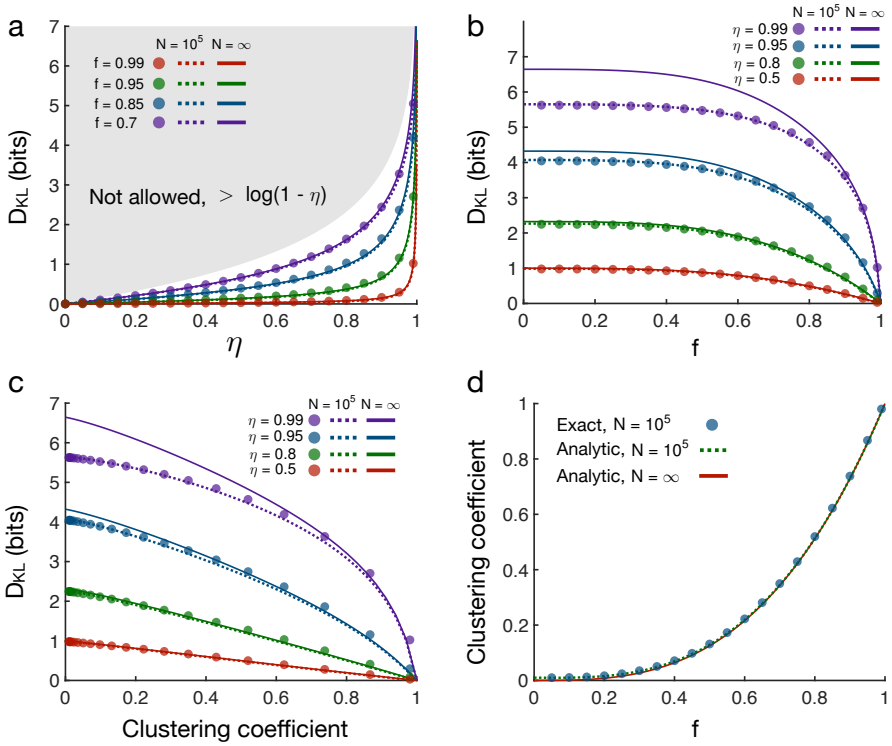


Figure 7.24: KL divergence from human expectations in stochastic block networks. (*a*) KL divergence as a function of the integration parameter η for stochastic block networks with average degree $\langle k\rangle = 100$ and communities of size $N_c = 100$. Data points are exact calculations for networks of size $N = 10^4$. Dashed lines are analytic predictions using Eq. (7.38) with $N = 10^4$. Solid lines are analytic results for the thermodynamic limit $N \to \infty$. (*b*) KL divergence as a function of the fraction of within-community edges f for different values of η. (*c*) KL divergence as a function of the average clustering coefficient for variable f and different values of η. (*d*) Average clustering coefficient as a function of f. Dashed line is the analytic prediction in Eq. (7.41) with $N = 10^4$. Solid line is the analytic result in the limit $N \to \infty$.

$p^{in} = f\langle k\rangle/N_c$ and the probability of an edge between two nodes in different communities as $p^{out} = (1-f)\langle k\rangle/(N-N_c)$. We then arrive at the following approximation,

$$\begin{aligned}\langle \triangle_i\rangle &= \frac{k_i^{in}(k_i^{in}-1)}{2}p^{in} + k_i^{in}k_i^{out}p^{out} + \frac{k_i^{out}(k_i^{out}-1)}{2}\left[\frac{N_c-1}{N-N_c-1}p^{in} + \left(\frac{N-2N_c}{N-N_c-1}\right)p^{out}\right] \\ &\approx \frac{p^{in}}{2}(k_i^{in})^2 + p^{out}k_i^{out}\left(k_i^{in} + \frac{k_i^{out}}{2}\right),\end{aligned} \tag{7.39}$$

where the approximation follows from the assumptions that $N \gg N_c$ and $k_i^{in}, k_i^{out} \gg 1$. Plugging in for p^{in}, p^{out}, $k_i^{in} = fk_i$, and $k_i^{out} = (1-f)k_i$, we have

$$\langle \triangle_i\rangle \approx \frac{\langle k\rangle k_i^2}{2}\left(\frac{f^3}{N_c} + \frac{(1+f)(1-f)^2}{N-N_c}\right). \tag{7.40}$$

Thus the average clustering coefficient is given by

$$\frac{1}{N}\sum_i \frac{2\langle \triangle_i \rangle}{k_i(k_i-1)} \approx \langle k \rangle \left(\frac{f^3}{N_c} + \frac{(1+f)(1-f)^2}{N-N_c} \right), \tag{7.41}$$

where the approximation follows from the assumption that $k_i \gg 1$. We see that this analytic result accurately predicts the increase in the average clustering coefficient with increasing modularity f (Fig. 7.24d). More importantly, we find that the KL divergence decreases with increasing clustering for fixed η and $\langle k \rangle$ (Fig. 7.24c). This final result indicates that increased modularity helps human observers maintain accurate representations, thereby reducing their inefficiency when processing information.

7.8.13 *Hierarchically modular networks*

The combination of high entropy and low KL divergence exhibited by real networks is driven by heterogeneous degrees and modular structure. Interestingly, degree heterogeneity and modularity are ubiquitous in natural and human-made systems (*50*, *51*, *121*, *207*, *250*, *461*, *570*), and together they define hierarchically modular organization (*547*). In order to simultaneously study entropy and KL divergence, it is helpful to have a model for generating networks with variable heterogeneity and modularity. One of the earliest models of hierarchical systems was developed to understand metabolic networks (*547*, *548*). Yet this model is deterministic, generating fractal networks in which it is difficult to tune the heterogeneity or modularity. Another common model is the nested stochastic block model (*25*, *26*), wherein small modules are nested inside larger modules. However, this model does not include heterogeneous degrees (a heavy-tailed degree distribution). Perhaps the closest model to what we require was recently developed to study the emergence of complex dynamics in the brain (*723*). In this model, the nested stochastic block model is combined with a preferential attachment rule to generate a rich club of hub nodes.

Here we propose a model that directly combines the static model (*128*, *258*, *392*) and the stochastic block model (*177*). Beginning with N disconnected nodes, we first assign each node i a weight $w_i = i^{-\alpha}$, where $\alpha \in [0, 1]$ is related to the scale-free exponent by $\gamma = 1 + \frac{1}{\alpha}$. We also assign each node i to a community. Then, we randomly select pairs of nodes i and j within the same community with probabilities proportional to their weights, and we connect them if they have not already been connected. This process is repeated until $fE = \frac{1}{2}f\langle k \rangle N$ edges have been added within communities. We then repeat this process again until $(1-f)E = \frac{1}{2}(1-f)\langle k \rangle N$ edges have been added between communities. The resulting network has a degree distribution that drops off as a power law $\mathcal{P}(k) \sim k^{-\gamma}$ and also has the same community structure as a stochastic block model.

Sweeping over the two parameters γ and f, while fixing the average degree $\langle k \rangle = 100$ and community size $N_c = 100$, we see that our hierarchically modular model exhibits a variety of entropies (Fig. 7.25a) and KL divergences (Fig. 7.25b). Additionally, we

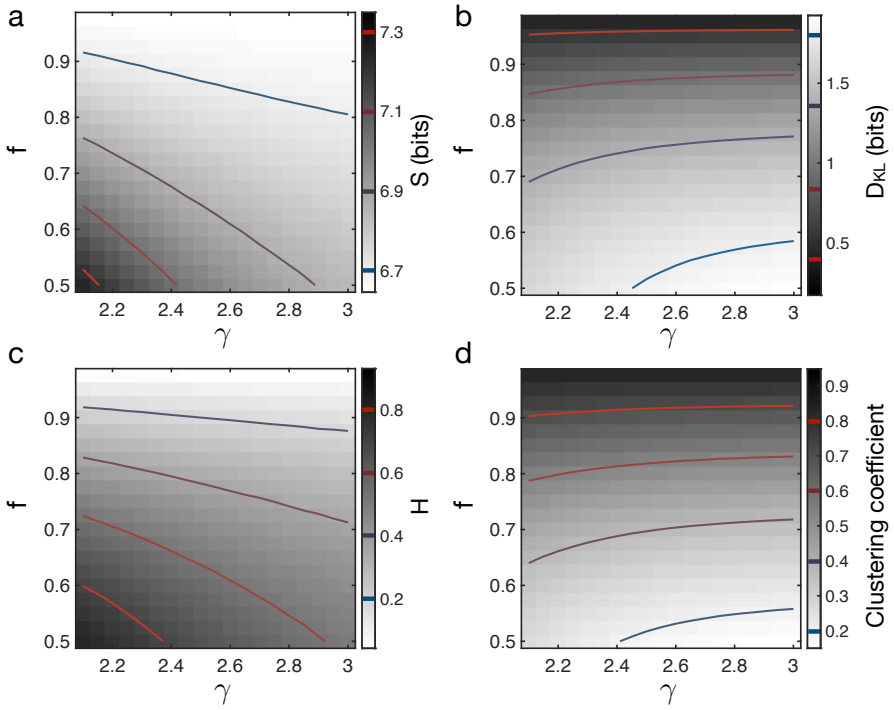


Figure 7.25: Information and structural properties of hierarchically modular networks. (*a*) Entropy as a function of the scale-free exponent γ and the fraction of within-community edges f for hierarchically modular networks with average degree $\langle k \rangle = 100$ and communities of size $N_c = 100$. Each point is an exact calculation for a network of size $N = 10^4$. (*b*) KL divergence as a function of γ and f in the same networks with η fixed to the average value 0.80 from our experiments. (*c*) Degree heterogeneity H varies as a function of γ and f in a similar fashion to the entropy (*a*). (*d*) Average clustering coefficient varies as a function of γ and f much like the KL divergence (*b*).

verify that the model can attain a wide range of degree heterogeneities (Fig. 7.25c) and clustering coefficients (Fig. 7.25d). Notably, the variation in the degree heterogeneity and clustering coefficient with γ and f appears almost identical to the variation in the entropy and KL divergence, respectively, once again indicating that entropy is primarily driven by heterogeneity and KL divergence is primarily driven by clustering or modularity.

Given our investigation of the information properties of different network models, it is ultimately important to compare against real communication networks. For each network listed in Tab. 7.13, we generate series of scale-free networks with various exponents γ, stochastic block networks with various within-community fractions f, and hierarchically modular networks with various exponents γ (for fixed f) and various f (for fixed γ). Each model network maintains the same number of nodes N and edges E as the corresponding real network. For the stochastic block and hierarchically modular networks, we choose community sizes that are roughly the square root of the network size $N_c \approx \sqrt{N}$ for the purpose of remaining consistent with our model-based analysis (wherein $N = 10^4$ and $N_c = \sqrt{10^4} = 100$). Comparing each real and model network

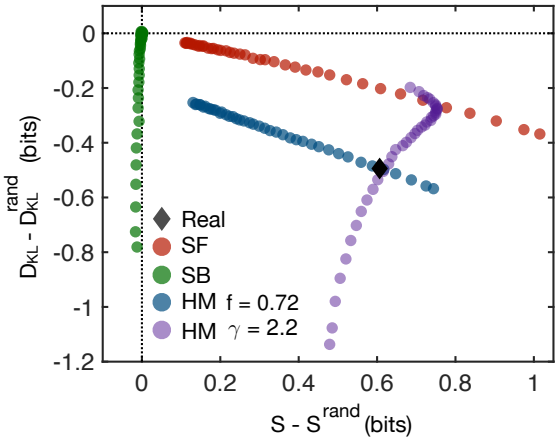


Figure 7.26: Comparing the information properties of real and model networks. Entropies and KL divergences of real and model networks compared to fully randomized versions. For each model network in Tab. 7.1, we generate SF networks with variable γ (red), SB networks with communities of size $N_c \approx \sqrt{N}$ and variable f (green), and HM networks with $N_c \approx \sqrt{N}$ and variable γ (fixed $f = 0.72$; blue) or variable f (fixed $\gamma = 2.2$; purple), all with the same number of nodes N and edges E as the real network. Each real and model network is then compared with 100 randomized versions; data points are first averaged over the 100 randomized networks and then averaged over the set of real networks in Tab. 7.1. HM networks with $\gamma = 2.2$ and $f = 0.72$ match the average entropy and KL divergence of real networks.

with completely randomized versions of the same networks (Fig. 7.26), we find that: (i) scale-free networks cannot attain the low KL divergence displayed by real networks and (ii) stochastic block networks cannot attain the high entropy displayed by real networks, but (iii) hierarchically modular networks can achieve both with a parameter combination of $\gamma \approx 2.2$ and $f \approx 0.72$. Thus, we confirm that both heterogeneous degrees and modular structure are required (that is, hierarchical organization is required) to match the information properties of real networks.

7.8.14 *Network datasets*

The real-world networks analyzed in the main text are listed and briefly described in Tab. 7.13. While the semantic, web, citation, and social networks are gathered from online network repositories, the language and music networks are novel. For the language networks, we developed code to (i) remove punctuation and white space, (ii) filter words by their part of speech, and (iii) record the transitions between the filtered words. Here we focus on networks of transitions between nouns, noting that the same methods could be used to record transitions between other parts of speech. The raw text was gathered from Project Gutenberg (`gutenberg.org/wiki/Main_Page`).

For the music networks, we read in audio files in MIDI format using the `readmidi` function in MATLAB (R2018a). For each song, we split the notes by their channel, which represents the different instruments. For each channel, we created a network of

note transitions. We then create a transition network representing the entire song by aggregating the transitions between notes across the different channels. The MIDI files were gathered from `midiworld.com` and from `kunstderfuge.com`. Our code and data are available upon request from the corresponding author.

Type	Name	N	E	Description
Language	Shakespeare*+ (*602*)	11,234	97,892	Noun transitions in Shakespeare's work.
	Homer*+ (*534*)	3,556	23,608	Same as above (Homer's Iliad).
	Plato*+ (*348*)	2,271	9,796	Same as above (Plato's Republic).
	Jane Austen*+ (*34*)	1,994	12,120	Same as above (Pride and Prejudice).
	William Blake*+ (*91*)	370	781	Same as above (Songs of Innocence...).
	Miguel de Cervantes*+ (*132*)	6,090	43,682	Same as above (Don Quixote).
	Walt Whitman*+ (*707*)	4,791	16,526	Same as above (Leaves of Grass).
Semantic	Bible (*383*)	1,707	9,059	Pronoun co-occurrences in Bible verses.
	Les Miserables (*383*)	77	254	Character co-occurrences.
	Edinburgh Thesaurus* (*64, 372*)	7,754	226,518	Word similarities in human experiments.
	Roget Thesaurus* (*64, 561*)	904	3,447	Linked semantic categories.
	Glossary terms (*64*)	60	114	Words used in definitions of other words.
	FOLDOC* (*64, 325*)	13,274	90,736	Same as above (computing terms).
	ODLIS* (*64, 553*)	1,802	12,378	Same as above (information science terms).
Web	Google internal* (*383, 499*)	12,354	142,296	Hyperlinks between Google's own cites.
	Education (*253, 568*)	2,622	6,065	Hyperlinks between education webpages.
	EPA (*172, 568*)	2,232	6,876	Pages linking to www.epa.gov.
	Indochina (*94, 568*)	9,638	45,886	Hyperlinks between pages in Indochina.
	2004 Election blogs* (*5, 383*)	793	13,484	Hyperlinks between blogs on US politics.
	Spam (*127, 568*)	3,796	36,404	Hyperlinks between spam pages.
	WebBase (*94, 568*)	6,843	16,374	Hyperlinks gathered by web crawler.
Citations	arXiv Hep-Ph*+ (*383, 397*)	12,711	139,500	Citations in Hep-Ph section of the arXiv.
	arXiv Hep-Th*+ (*383, 397*)	7,464	115,932	Citations in Hep-Th section of the arXiv.
	Cora* (*383, 643*)	3,991	16,621	Citation network between scientific papers.
	DBLP* (*383, 401*)	240	858	Citation network between scientific papers.
Social	Facebook+ (*383, 687*)	13,130	75,562	Subset of the Facebook network.
	arXiv Astr-Ph (*383, 397*)	17,903	196,972	Coauthorships in Astr-Ph section of arXiv.
	Adolescent health* (*383, 455*)	2,155	8,970	Friendships between students.
	Highschool* (*152, 383*)	67	267	Friendships between highschool students.
	Jazz (*254, 383*)	198	2,742	Collaborations between jazz musicians.
	Karate club (*383, 721*)	34	78	Interactions between karate club members.
Music	Thriller – Michael Jackson*+ (*336*)	67	446	Network of note transitions.
	Hard Day's Night – Beatles*+ (*660*)	41	212	Same as above.
	Bohemian Rhapsody – Queen*+ (*541*)	71	961	Same as above.
	Africa – Toto*+ (*670*)	39	163	Same as above.
	Sonata No 11 – Mozart*+ (*463*)	55	354	Same as above.
	Sonata No 23 – Beethoven*+ (*73*)	69	900	Same as above.
	Nocturne Op 9-2 – Chopin*+ (*144*)	59	303	Same as above.
	Clavier Fugue 13 – Bach*+ (*40*)	40	143	Same as above.
	Ballade Op 10-1 – Brahms*+ (*102*)	69	670	Same as above.

Table 7.13: Real networks analyzed in the main text. For each network we list its type; name, reference, whether it has a directed version (denoted by $*$), and whether it has a temporally evolving version (denoted by $+$); number of nodes N; number of edges E; and a brief description.

Part III

THE STATISTICAL PHYSICS OF NEURAL DYNAMICS

In Parts I and II, we used foundational ideas from statistical mechanics, including the maximum entropy principle, the fluctuation-dissipation theorem, and the free energy principle, to describe how collective activity emerges in human populations and to uncover how individual humans learn and process information using complex networks. Both collective activity and individual behavior, however, fundamentally arise from the correlated firing of billions of neurons at the scale below. In Chapter 8, we provide an overview of current efforts to understand the brain's complex dynamics that draw on intuitions, models, and theories from physics, spanning the domains of statistical mechanics, information theory, and dynamical systems and control. For example, recent advances in non-equilibrium statistical mechanics have revealed that enzymatic and metabolic processes drive the brain away from equilibrium at small scales. Yet it remains unclear if an how non-equilibrium dynamics manifest at macroscopic scales. In Chapter 9, we present a framework to probe for non-equilibrium dynamics in complex living systems. We apply our method to whole-brain neuroimaging data, demonstrating not only that the brain operates out of equilibrium, but that it functions farther from equilibrium during periods of physical and cognitive exertion. Together, these results establish that non-equilibrium dynamics can arise at macroscopic scales and provide a general tool for quantifying the non-equilibrium nature of complex systems.

8

THE PHYSICS OF BRAIN NETWORK STRUCTURE, FUNCTION, AND CONTROL

This chapter contains work from Lynn, Christopher W., and Danielle S. Bassett. "The physics of brain network structure, function and control." Nature Reviews Physics 1.5 (2019): 318.

Abstract

The brain is a complex organ characterized by heterogeneous patterns of structural connections supporting unparalleled feats of cognition and a wide range of behaviors. New noninvasive imaging techniques now allow these patterns to be carefully and comprehensively mapped in individual humans and animals. Yet, it remains a fundamental challenge to understand how the brain's structural wiring supports cognitive processes, with major implications for the personalized treatment of mental health disorders. Here, we review recent efforts to meet this challenge that draw on intuitions, models, and theories from physics, spanning the domains of statistical mechanics, information theory, and dynamical systems and control. We begin by considering the organizing principles of brain network architecture instantiated in structural wiring under constraints of symmetry, spatial embedding, and energy minimization. We next consider models of brain network function that stipulate how neural activity propagates along these structural connections, producing the long-range interactions and collective dynamics that support a rich repertoire of system functions. Finally, we consider perturbative experiments and models for brain network control, which leverage the physics of signal transmission along structural wires to infer intrinsic control processes that support goal-directed behavior and to inform stimulation-based therapies for neurological disease and psychiatric disorders. Throughout, we highlight several open questions in the physics of brain network structure, function, and control that will require creative efforts from physicists willing to brave the complexities of living matter.

8.1 INTRODUCTION

It is our good fortune as physicists to seek to understand the nature of the observable world around us. In this inquiry, we need not reach to contemporary science to appreciate the fact that our perception of the world around us is inextricably linked to the world within us: the mind. Indeed, even Aristotle c. 350 B.C. noted that it is by mapping the structure of the world that the human comes to understand their own

mind (*391*). "Mind thinks itself because it shares the nature of the object of thought; for it becomes an object of thought in coming into contact with and thinking its objects, so that mind and object of thought are the same" (*27*). Over the ensuing 2000-plus years, it has not completely escaped notice that the mappers of the world have unique contributions to offer the mapping of the mind (from Thales of Miletus, c. 624–546 B.C., to Leonardo Da Vinci, 1452–1519). More recently, it is notable that nearly all famous physicists of the early 20th century – Albert Einstein, Niels Bohr, Erwin Schroedinger, Werner Heisenberg, Max Born – considered the philosophical implications of their observations and theories (*632*). In the post-war era, philosophical musings turned to particularly conspicuous empirical contributions at the intersection of neuroscience and artificial intelligence, spanning polymath John von Neumann's work enhancing our understanding of computational architectures (*690*) and physicist John Hopfield's invention of the associative neural network, which revolutionized our understanding of collective computation (*322*).

In the contemporary study of the mind and its fundamental organ – the brain – nearly all of the domains of physics, perhaps with the exception of relativity, are not only relevant but truly essential, motivating the early coinage of the term *neurophysics* some four decades ago (*596*). The fundamentals of electricity and magnetism prove critical for building theoretical models of neurons and the transmission of action potentials (*378*). These theories are being increasingly informed by mechanics to understand how force-generating and load-bearing proteins bend, curl, kink, buckle, constrict, and stretch to mediate neuronal signaling and plasticity (*675*). Principles from thermodynamics come into play when predicting how the brain samples the environment (action) or shifts the distribution of information that it encodes (perception) (*232*). Collectively, theories of brain function are either buttressed or dismantled by imaging, with common tools including magnetic resonance imaging (*530*) and magnetoencephalography (*294*), the latter being built on superconducting quantum interference devices and next-generation quantum sensors that can be embedded into a system that can be worn like a helmet, revolutionizing our ability to measure brain function while allowing free and natural movement (*98*). Moreover, recent developments in nanoscale analysis tools and in the design and synthesis of nanomaterials have generated optical, electrical, and chemical methods to explore brain function by enabling simultaneous measurement and manipulation of the activity of thousands or even millions of neurons (*14*). Beyond its relevance for continued imaging advancements (*525*), optics has come to the fore of neuroscience over the last decade with the development of optogenetics, an approach that uses light to alter neural processing at the level of single spikes and synaptic events, offering reliable, millisecond-timescale control of excitatory and inhibitory synaptic transmission (*100*).

Such astounding advances, enabled by the intersection of physics and neuroscience, have motivated the construction of a National Brain Observatory at the Argonne National Laboratory (Director: Peter Littlewood, previously of Cavendish Laboratories) funded by the National Science Foundation, as well as frequent media coverage including titles in the APS News such as "Physicists, the Brain is Calling You."(*535*)

And as physicists answer the call, our understanding of the brain deepens and our ability to mark and measure its component parts expands. Yet alongside this growing systematization and archivation, we have begun to face an increasing realization that it is the interactions between hundreds or thousands of neurons that generate the mind's functional states (*14*). Indeed, from interactions among neural components emerge computation (*435*), communication (*228*), and information propagation (*82*). We can confidently state of neuroscience what Henri Poincare, the French mathematician, theoretical physicist, and philosopher of science, states of science generally: "The aim of science is not things themselves, as the dogmatists in their simplicity imagine, but the relations among things; outside these relations there is no reality knowable." (*531*) The overarching goal of mapping these interactions in neural systems has motivated multibillion-dollar investments across the United States (the Brain Initiative generally, and the Human Connectome Project specifically (*678*)), the European Union (the Blue Brain Project (*426*)), China (the China Brain Project (*533*)), and Japan (Japan's Brain/MINDS project (*489*)).

While it is clear that interactions are paramount, exactly how the functions of the mind arise from these interactions remains one of the fundamental open questions of brain science (*59*). To the physicist, such a question appears to exist naturally within the purview of statistical mechanics (*600*), with one major caveat: the interaction patterns observed in the brain are far from regular, such as those observed in crystalline structures, and are also far from random, such as those observed in fully disordered systems (*57*). Indeed, the observed heterogeneity of interaction patterns in neural systems – across a range of spatial and temporal scales – generally limits the utility of basic continuum models or mean-field theories, which would otherwise comprise our natural first approaches. Fortunately, similar observations of interaction heterogeneity have been made in other technological, social, and biological systems, leading to concerted efforts to develop a statistical mechanics of complex networks (*11*). The resultant area of inquiry includes criteria for building a network model of a complex system (*118*), statistics to quantify the architecture of that network (*159*), models to stipulate the dynamics that can occur both in and on a network (*283*, *289*, *727*), and theories of network function and control (*460*, *661*).

Here, we provide a brief review for the curious physicist, spanning the network-based approaches, statistics, models, and theories that have recently been used to understand the brain. Importantly, the interpretations that can be rationally drawn from all such efforts depend upon the nature of the network representation (*118*), including its descriptive, explanatory, and predictive validity – topics that are treated with some philosophical rigor elsewhere (*63*). Following a simple primer on the nature of network models, we discuss the physics of brain network structure, beginning with an exposition regarding measurement before turning to an exposition regarding modeling. In a parallel line of discourse, we then discuss the physics of brain network function, followed by a description of perturbation experiments and brain network control. In each section we separate our remarks into the known and the unknown, the past and the future, the fact and the speculation. Our goal is to provide an

accessible introduction to the field, and to inspire the younger generation of physicists to courageously tackle some of the most pressing open questions surrounding the inner workings of the mind.

8.2 THE PHYSICS OF BRAIN NETWORK STRUCTURE

We begin with a discussion of the architecture, or structural wiring, of networks in the brain, focusing on the measurement and modeling of their key organizational features (see Fig. 8.1 for a simple primer on networks). Each edge in a structural brain network represents a physical connection between two elements. For example, synapses support the propagation of information between neurons (*517*) and white matter tracts define physical pathways of communication between brain regions (*36*). In physics, it has long been recognized that the organization of such structural connections can determine the qualitative large-scale features of a system (*11*). In the Ising model, for instance, a one-dimensional lattice remains paramagnetic across all temperatures (*334*), while in two dimensions or more, the system spontaneously breaks symmetry, yielding the type of bulk magnetization exhibited by magnets on a refrigerator (*113*, *491*). Similarly, the organization of structural wiring in the brain largely determines the types of mental processes and cognitive functions that can be supported (*438*, *449*, *520*, *621*, *623*), from memory (*136*, *545*, *693*) to learning (*307*, *659*), and from vision (*649*) to motion (*728*). However, unlike many physics applications, which assume simple lattice or random network architectures, the wiring of the brain is highly heterogeneous, often making symmetry arguments and mean-field descriptions far from applicable (*57*). While this heterogeneity presents a unique set of challenges, in what follows we review some powerful experimental and theoretical tools that allow us to distill the brain's structural complexity to a number of fundamental organizing principles.

8.2.1 *Measuring brain network structure*

Some of the earliest empirical measurements of the brain's structural connectivity can be traced to Camillo Golgi, who in 1873 soaked blocks of brain tissue in silver-nitrate solution to provide among the first glimpses of the intricate branching of nerve cells (*261*). Soon after, Santiago Ramón y Cajal combined Golgi's method with light microscopy to achieve stunning pictures establishing that neurons do not exist in solitude; they instead combine to form intricate networks of physical connections (*714*). This notion that the brain comprises a complex web of distinct components, known as the neuron doctrine (*605*), established the foundation upon which modern network neuroscience has flourished. The introduction of the electron microscope in the 1930s provided even more detailed measurements of the physical connections between neurons. Perhaps the most impressive application remains the complete mapping of interconnections between the 302 neurons in the nematode *C. elegans* (*705*). Since this achievement, reconstructions of the synaptic connectivity in other animals have

A simple primer on networks. Here, we define what we mean by a network and describe tools for summarizing its architecture. Importantly, a network is agnostic to the system that it represents (*63*), whether it be a brain, a granular material (*505*), or a quantum lattice (*90*). By far the simplest network model is represented by a binary undirected graph in which identical nodes represent system components and identical edges indicate relations or connections between pairs of nodes (see the figure). Such a network can be encoded in an adjacency matrix $\mathbf{A}$, where each element A_{ij} indicates the strength of connectivity between nodes i and j. When all edge strengths are unity, the network is said to be binary. When edges have a range of weights, the network represented by the adjacency matrix is said to be weighted. When $\mathbf{A} = \mathbf{A}^T$, the network is undirected; otherwise, the network is directed.

One can extend this simple encoding to study multilayer, multislice, and multiplex networks (*373*); dynamic or temporal networks (*171*, *317*); annotated networks (*474*); hypergraphs (*62*); and simplicial complexes (*252*). One can also calculate various statistics to quantify the architecture of a network and to infer the function thereof (see figure). Intuitively, these statistics range from measures of the local structure in the network, which depend solely on the links directly emanating from a given node (e.g., degree and clustering), to measures of the network's global structure, which depend on the complex pattern of interconnections between all nodes (e.g., path lengths and centrality) (*159*). Intermediate statistics exist to study network organization at the mesoscale, such as cavity structure and community structure, the latter of which describes the presence of communities of densely connected nodes (*223*, *224*, *536*). As we will see, the encoding of a system as a network and the quantitative assessment of its architecture can provide important insights into its function (*661*, *699*).

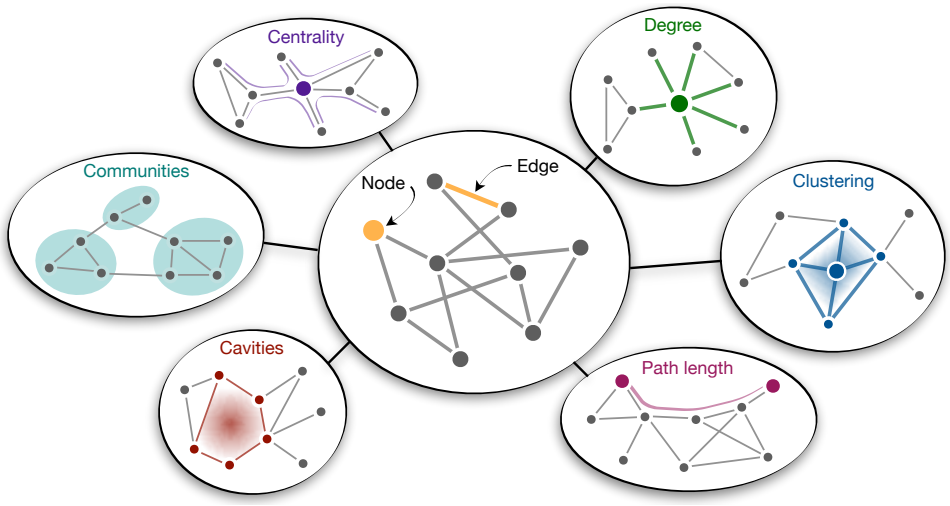


Figure 8.1: A primer on network properties. (*Center*) Nodes, illustrated by circles, represent stimuli, items, or states in a sequence. Edges, illustrated by lines, connect pairs of nodes if it is possible to transition from one node to the other. The organization of edges among nodes is referred to as the network's *topology* or *structure*. (*Circumjacent*) A network's topology can be described using properties that characterize its local, mesoscale, or global organization.

evolved rapidly, from a mapping of the optic medulla in the visual system of the fruit fly *Drosophila* to the enumeration of connections between 950 distinct neurons in the mouse retina (*304*, *649*). Efforts continue to press forward toward the ultimate goal of reconstructing the neuronal wiring diagram of an entire human brain (*627*).

Concurrently with these achievements using electron microscopy, complimentary efforts in tract tracing have revealed the mesoscale structure of the macaque (*425*, *633*), cat (*719*), mouse (*488*), and fly (*609*). Particularly important for our understanding of human cognition are recent advances in noninvasive imaging that have allowed unprecedented views of the mesoscale structure of the brain *in vivo*. Introduced in the 1970s, computerized axial tomography (CAT) provided among the most detailed anatomic images of the human brain to date (*327*). Soon after, the development of magnetic resonance imagining (MRI) sparked an explosion of refinements, a notable example being diffusion tensor imaging (DTI) (*526*). While standard CAT and MRI techniques capture cross-sectional images of the brain, DTI traces the diffusion of water molecules through white matter tracts to reconstruct the large-scale neural pathways connecting distinct brain regions (*56*, *75*). Given measurements of the anatomical wiring connecting a set of neural elements, such as synapses linking neurons or white matter tracts connecting brain regions, researchers can build a structural brain network by forming edges between elements that share a physical connection (Fig. 8.2a). Ongoing experimental efforts to acquire these measurements continue to provide rich network datasets detailing the brain's structural organization.

8.2.2 *Modeling brain network structure*

A first glance at the brain's wiring reveals that it is far from homogeneous – a fact that is not surprising considering the array of physical, energetic, and cognitive constraints that it is required to balance (*116*). To handle this heterogeneity, researchers have increasingly turned to the field of network science for mathematical tools and intuitions (*60*, *80*). The primary goal of this interdisciplinary effort has been to distill the explosion of experimental data, spanning structural brain networks in *C. elegans* (*481*), the mouse (*305*), cat (*86*), macaque (*87*, *203*), and human (*84*), down to a number of cogent organizing principles. Here we review some important properties that are thought to characterize structural brain networks and introduce several generative network models that help to explain how these properties arise from underlying biological mechanisms (Fig. 8.2b).

8.2.2.1 *Random structure*

While healthy members of a species exhibit anatomical similarities in brain structure, the specific instantiation of physical connections in each individual is far from deterministic. Indeed, *in vivo* imaging techniques in humans, such as DTI described above, have revealed not only stark differences in brain structure between individuals (*663*), but also within the same individual over time (*269*, *549*). Importantly, these structural

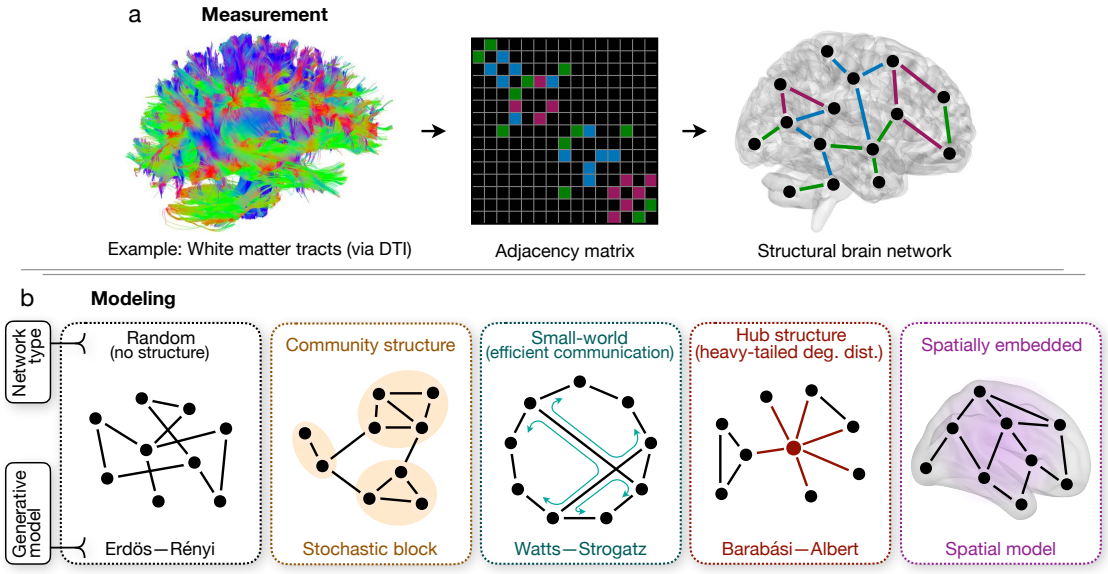


Figure 8.2: Measuring and modeling brain network structure. (*a*) The measurement of brain network structure begins with experimental data specifying the physical interconnections between neurons or brain regions. As an example, we consider a dataset of white matter tracts measured via DTI. First, the data is discretized into non-overlapping gray matter volumes representing distinct nodes. Then, one constructs an adjacency matrix $\mathbf{A}$, where A_{ij} represents the connection strength between nodes i and j. This adjacency matrix, in turn, defines a structural brain network constructed from our original measurements of physical connectivity. (*b*) To capture an architectural feature of structural brain networks, we utilize generative network models. The simplest generative network model is the Erdös–Rényi model, which has no discernible non-random structure. Networks with modular structure, divided into communities with dense connectivity, are constructed using the stochastic block model. Small-world networks, which balance efficient communication and high clustering, are generated using the Watts–Strogatz model. Networks with hub structure, characterized by a heavy-tailed degree distribution, are typically constructed using a preferential attachment model such as the Barabási–Albert model. Spatially embedded networks, whose connectivity is constrained to exist within a physical volume, are generated through the use of spatial network models.

differences have been linked to variability in a wide range of behaviors (*355*), including empathy (*47*), introspection (*219*), fear acquisition (*296*), and even political orientation (*356*). To study the mathematical properties of random networks, and to understand the types of biological mechanisms that can give rise to qualitative structural properties, it is useful to consider generative network models (*80*). The simplest and most common model for generating random networks is the Erdös–Rényi (ER) model (*205*), wherein each pair of nodes is connected independently with a fixed probability P. While the ER model has a number of interesting mathematical properties, such as a binomial degree distribution, it has no discernible structure and does not reflect the mechanisms by which most networks grow in the brain. Accordingly, if we wish to understand some of the principles underlying naturally occurring brain networks, we must consider generative models that yield networks with realistic properties.

8.2.2.2 *Community structure*

Perhaps the brain's most well-studied structural property is its division into distinct anatomical regions, which are widely thought to be responsible for specialized cognitive functions (*606*). Interestingly, by studying the large-scale structure of brain networks in several mammalian species, researchers have shown that the organization of connections tends to partition the networks into densely-connected communities separated by sparse inter-community connectivity (*310*, *624–626*). Moreover, these clusters of high connectivity closely resemble postulated anatomical subdivisions (*310*). It has therefore been argued that the so-called community structure of brain networks segregates the brain into subnetworks with specific cognitive functions (*38*, *61*, *394*, *618*, *656*). Practically speaking, in order to extract the community structure of a real-world network, one must employ algorithms for community detection – a vibrant branch of research that is now applied throughout network neuroscience (*81*, *366*). From a complimentary perspective, to generate networks with a defined community structure, researchers predominantly use the stochastic block (SB) model, wherein nodes are assigned to distinct communities and an edge is placed between each pair of nodes with a probability that depends on the nodes' community assignments (*10*, *83*). Such SB networks are often used as null models to distinguish between properties of brain networks that are implied simply by their community structure and those that require additional biological mechanisms (*80*, *83*).

8.2.2.3 *Small-world structure*

Seemingly in contradiction to their striking community structure, large-scale brain networks also exhibit average path lengths between all nodes that are much shorter than a typical random network (*116*, *404*, *681*). This competition between high clustering and short average paths is thought to facilitate the simultaneous segregation and integration of information in the brain (*179*), possibly minimizing the total number of computational steps needed to process external stimuli (*354*, *390*). Seeking an explanation for similar "small-world" topologies exhibited by other real-world systems (most notably social networks (*671*)), Duncan Watts and Steven Strogatz developed a model for generating random networks with both high clustering and short average path lengths (*699*). Generally, the Watts-Strogatz (WS) model supposes that small-world networks are an interpolation between two extreme configurations: a ring lattice, wherein nodes are arranged along a circle and connected to their k nearest neighbors on either side, and an ER random network. Notably, the presence of small-world structure in the brain suggests that efficient communication emerges from a finely-tuned balance of lattice-like organization and structural disorder.

8.2.2.4 *Hub structure*

In addition to their modular and small-world structure, many large-scale brain networks also feature high-degree "hubs", which form a densely interconnected structural

core (*268*). Acting as bridges between structurally distinct communities, these specialized hub regions are thought to help minimize overall path lengths across the network (*625*) and facilitate the integration of information (*179*). Supporting the notion of a centralized core, many studies have identified hubs within the parietal and prefrontal regions, areas that are often active during a wide range of cognitive functions (*268*, *700*). Such core-periphery architecture is characterized by a heavy-tailed degree distribution, such as that observed in scale-free networks, in some cases arising through preferential attachment mechanisms (*175*). In the Barabàsi–Albert (BA) model (*50*), for instance, nodes are added to a network in sequential order, and each new node i forms an edge with each existing node j with a probability proportional to the degree of node j. In this way, new nodes preferentially attach to existing nodes of high degree, creating a "rich club" of centralized hubs that link otherwise distant regions of the network.

8.2.2.5 *Spatial structure*

Thus far, we have focused exclusively on the topological properties of brain networks, which are thought to be driven primarily by the simultaneous functional pressures of information segregation and integration (*179*). However, brain networks are also physically constrained to exist within a tight three-dimensional volume and their structural connections are metabolically driven to minimize total wiring distance (*61*, *116*, *354*). Such physical and metabolic constraints are captured by spatial (or geometric) network models, which embed networks into three-dimensional Euclidean space and penalize the formation of long-distance connections (*80*). The simplest such model assumes that the probability of two nodes i and j forming an edge is proportional to $d_{ij}^{-\alpha}$, where d_{ij} is the physical distance between i and j, and $\alpha \geqslant 0$ tunes the metabolic cost associated with constructing connections of a given length (*167*). If we keep the number of nodes and edges fixed, one can see that, much like the WS model, this spatial model interpolates between a lattice-like structure, in which nodes only connect to their nearest neighbors ($\alpha \to \infty$), and an ER random network ($\alpha = 0$).

8.2.2.6 *Competition between structural properties*

As the brain grows and adapts to changing cognitive demands, it is widely thought that the underlying network evolves to balance the trade-off between topological value and metabolic wiring cost (*116*). Thus, while the modular, small-world, heavy-tailed, and inherently physical properties of brain networks provide simple organizing principles, in reality the brain is constantly and dynamically weighing these pressures against one another. Accordingly, an accurate generative model should aim to explain multiple real-world properties at once (*80*). With this goal in mind, recent work has shown that an impressive range of topological properties can be understood as arising from a competition between two competing factors: a metabolic penalty for the formation of long-distance connections and a topological incentive to connect regions with similar inputs (*684*). Notably, investigations of the human, *C. elegans*, and mouse connectomes have revealed that the total wiring distance is consistently greater than

minimal, supporting the notion that brain networks weigh the costs of long-distance connections against the functional benefits of an integrated network topology (*61*, *572*). Together, these efforts toward a comprehensive generative model are vital for our understanding of healthy brain network structure, with important clinical implications for the diagnosis, prognosis, prevention, and treatment of disorders of mental health (*353*, *630*).

8.2.3 *The future of brain network structure*

Current advances in neuroimaging techniques and network science continue to expand our ability to measure and model the architecture of structural connections in the brain. As experimental measurements become increasingly detailed, an important direction is the bridging of brain network structure at different spatiotemporal scales (*106*, *137*, *591*). Such cross-scale approaches could link protein interaction networks within neurons to the wiring of synaptic connectivity between neurons to mesoscale networks connecting brain regions and all the way to social networks linking distinct organisms (Fig. 8.3). The goal of such cross-scale integration is to understand how the architecture of connectivity at each of these scales emerges from the scale below. Practically, researchers have begun to address this goal by employing hierarchical network models (*76*), which treat each node at the macroscale as an entire subnetwork at the microscale (*442*).

Perhaps the most ambitious future goal is the reconstruction of the entire human connectome at the scale of individual neurons, pressing the current boundaries of 3D electron microscopy and statistical image reconstruction (*627*). Extensive mapping efforts in other species have revealed notable and quantifiable neuronal diversity (*28*, *601*), suggesting the importance of extending network models to include non-identical units. At the mesoscale, advances in noninvasive imaging have allowed researchers to begin tracking changes in structural connectivity over time (*68*, *481*, *592*, *730*). To analyze these temporally ordered measurements, network scientists have extended standard static graph theoretic tools to study networks with dynamically evolving connections (*366*). Notably, these so-called temporal networks (*317*) were recently shown to be easier to control, requiring less energy to attain a desired pattern of neural activity, than their static counterparts (*403*).

Properly modeling the dynamics *of* brain networks requires also understanding the functional dynamics occurring *on* brain networks. For instance, dating to Donald Hebb's 1949 book *The Organization of Behavior*, it has been posited that the strength of a synaptic connection increases with the persistent synchronized firing of its pre- and postsynaptic neurons (*301*). Such Hebbian plasticity has been observed *in vitro* (*422*) and is thought to explain many aspects of brain network structure (*453*, *619*). More generally, Hebb's postulate highlights the fact that a complete understanding of the brain cannot simply include a description of its structural wiring; it must also stipulate the types of dynamics supported by this physical circuitry.

Bridging spatiotemporal scales. In the context of complex systems generally and neural systems specifically, the cutting edge work relates to extending our tools, theories, and intuitions from a single network to so-called multiscale, multilayer, and multiplex networks (*273*, *373*). Perhaps the most obvious context in which to make this extension is from regional networks to cellular-scale neuronal networks (*442*). Large-scale brain activity provides a coarse-grained encoding of neural processes, and the map from cellular dynamics to regional dynamics reflects rules of system function. By combining these two layers we can address questions like, "How do cellular processes shape circuit behavior?" The next logical extension is to move even further down the natural hierarchy of scales to understand how molecular networks – including gene coexpression networks (*28*, *557*, *565*, *704*) – shape the behavior of cells (*293*). Understanding how molecular mechanisms affect large-scale brain network function is critical for the development of effective pharmacological interventions (*104*, *413*, *630*). By extending the network model from regions to cells to molecular drivers, we can ask questions like, "How do genetic codes and epigenetic drivers shape circuit behavior across spatial scales?" And in a final extension, it is time to move up in the natural hierarchy of scales to combine information from the connectivity within a single human brain to the connectivity between human brains in large-scale social networks (*191*, *508*, *509*, *587*). While brain activity and structure offer biological mechanisms for human behaviors, social networks offer external inducers or modulators of those behaviors (*212*). By extending the network model to this larger scale, we can start to ask – and potentially answer – questions like, "How do brains shape social networks? And how do social ties shape the brain?" This extension will be important in understanding human behavior within the broader contexts of culture and society.

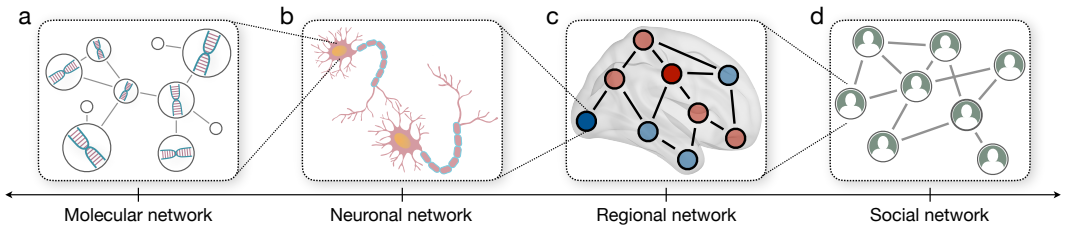


Figure 8.3: Brain networks at various scales. (*a*) Molecular networks composed of interacting molecules. (*b*) Neuronal networks composed of firing neurons. (*c*) Regional network composed of disparate brain areas communicating with one another. (*d*) Social network composed of individuals interacting with one another.

8.3 THE PHYSICS OF BRAIN NETWORK FUNCTION

While structural brain networks represent the physical wiring between neural elements (e.g., between individual neurons or brain regions), knowledge of this circuitry alone is not sufficient to understand how the brain *works*. For this reason, we turn our attention to models of brain network function that stipulate how neural activity propagates along structural connections. Just as the neuron doctrine postulates that the brain's structure is divided into a network of distinct nerve cells, it is also widely expected that the brain's array of cognitive functions emerges from the collective activity of individual neurons (*14*, *59*, *141*, *228*, *668*). To understand how the firing of simple nerve cells can give rise to the brain's rich repertoire of cognitive functions (*1*), analogies are often drawn with notions of emergence in statistical mechanics (*59*, *141*, *178*). Developed concurrently with the neuron doctrine in the late 19th century, statistical mechanics established (among other achievements) that the thermodynamic laws governing the macroscopic behavior of gas molecules can be derived from the microscopic dynamics of the molecules themselves (*551*). Similarly, growing evidence suggests that the dynamics of individual neurons and brain regions, when embedded in networks of structural connections, can produce the types of long-range correlations and collective patterns of activity that we observe in the brain (*109*, *110*, *141*, *398*, *589*, *622*, *692*). Here we traverse what is known about brain network function in relatively broad strokes, from the dynamics of distinct neurons to the networked activity of the entire brain.

8.3.1 *Measuring brain network function*

The first measurements of the brain's functional organization date to 1815, when Marie-Jean-Pierre Flourens pioneered the use of localized lesions in the brains of living animals to observe their effects on behavior. Through his experiments, Flourens discovered that the cerebellum regulates motor control, the cerebral cortex supports higher cognition, and the brain stem controls vital functions (*220*). The remainder of the 19th century brought increasingly detailed measurements of the brain's functional organization, from the demonstration that the occipital lobe regulates vision (*503*) to the discovery that the left frontal lobe is essential for speech (*108*). These discoveries, combined with the early images of neural circuits captured by Ramón y Cajal (*714*), culminated in Thomas Scott Sherrington's book *The Integrative Action of the Nervous System*, which proposed the idea that neurons behave in functional groups (*606*).

Meanwhile, in 1849 the physicist Hermann von Helmholtz achieved the first electrical measurements of a nerve impulse (*689*), sparking a wave of experiments investigating the electrical properties of the nervous system. Through invasive measurements in animals using newly-developed electroencephalography (EEG) techniques (*288*), it quickly became clear that individual neurons communicate with one another via electrical signals (*70*, *129*, *411*), thus providing a clear mechanism explaining how information is propagated and manipulated in the brain. Today, scientists possess a rich menu of experimental techniques for measuring brain dynamics across a range

of scales. At the neuronal level, the development of invasive methods in animals, such as electrophysiological recordings of brain slice preparations *in vitro* (*199*, *276*) and calcium imaging of neuronal activity *in vivo* (*278*, *638*), have vastly expanded our understanding of synaptic communication. At the regional level, complimentary minimally-invasive imaging techniques have identified fundamental properties of information processing in humans (*515*). Interestingly, these advances in mesoscale functional imaging can largely be traced to the efforts of physicists. MEG methods, for instance, use superconducting quantum interference devices (SQUIDS) to directly measure the magnetic fields generated by electrical currents in the brain (*98*, *292*); and PET techniques measure the positron emission of radioisotopes produced in cyclotrons to reconstruct the metabolic activity of neural tissue (*43*). Over the last twenty years, measurements of brain dynamics have been increasingly dominated by functional MRI (fMRI) (*544*), which estimates neural activity by calculating contrasts in blood oxygen levels, without relying on the invasive injections and radiation that limit the applicability of other imaging techniques (*724*). This modern progress in functional brain imaging has galvanized the field of network neuroscience by making detailed datasets of large-scale neural activity widely accessible.

One particularly important application of functional brain imaging has been the study of so-called functional brain networks (*677*), which have allowed researchers to investigate the organization of neural activity using tools from network science. In functional brain networks, as in their structural counterparts, nodes represent physical neural elements, ranging in size from individual neurons to distinct brain regions (*115*). However, whereas structural brain networks define the connectivity between elements based on physical measures of neural wiring (e.g., synapses between neurons or white matter tracts between brain regions), functional brain networks define connectivity based on the similarity between two elements' dynamics (*115*). To see how this works, we briefly consider the common example of a large-scale functional brain network calculated from fMRI measurements of regional activity (*677*) (Fig. 8.4a). First, blood oxygen levels indirectly reflecting neural activity are measured within three-dimensional non-overlapping voxels, spatially contiguous collections of which each represent a distinct brain region. After preprocessing the signal to correct for sources of systematic noise such as fluctuations in heart rate, the activity of each brain region is discretized in time, yielding a vector (or time series) of neural activity. Finally, to quantify functional connectivity, one computes the similarity between each pair of brain regions, for example using the quite simple Pearson correlation between the two regions' activity time series (*109*, *722*). The end result, even for different types of functional data and different choices for the preprocessing steps and similarity metric, is a functional brain network representing the organization of neural activity.

After constructing a functional brain network, researchers can utilize techniques from network science to study its key organizing features. Such efforts have demonstrated that large-scale functional brain networks, much like structural networks, exhibit signs of modular, small-world, heavy-tailed, and metabolically constrained organization (*2*, *79*, *299*, *577*, *677*). The existence of strong functional community structure, for instance,

further supports the hypothesis that brain networks segregate into subnetworks with specialized cognitive functions (*575*, *720*). Moreover, the presence of high clustering and short average path lengths, combined with the existence of high-degree hub regions, highlights the competing functional pressures of information segregation and integration in the brain (*2*, *58*). Metabolic constraints on the brain's structural wiring are also evident in its functional connectivity (*566*), with spatially localized brain regions generally supporting more strongly correlated activity than distant regions (*116*). In light of the similarities between the brain's functional and structural organization, it is tempting to suspect that functional brain networks closely resemble the physical wiring upon which they exist (*270*, *321*). However, the relationship between brain function and structure is highly nonlinear (*439*), and understanding how a functional brain network arises from its underlying structural connectivity remains a subject of intense academic focus (*106*, *507*).

8.3.2 *Modeling brain network function*

To understand how the web of physical connections in the brain gives rise to its functional properties, statistical mechanical intuition dictates that we should begin by studying the dynamics of individual elements. Once we have settled on accurate models of the interactions between individual neurons and brain regions, we can link these elements together in a network to predict macroscopic features of the brain's function from its underlying structure (*60*, *63*). Interestingly, the history of modeling in neuroscience has followed precisely this path, beginning with models of neuronal dynamics (*218*, *316*, *435*), then increasing in scale to mean-field neural mass models of distinct brain regions (*88*, *710*), and eventually achieving models of entire networks of neurons and brain regions (*322*, *384*, *589*). Here we review important developments in the modeling of neural dynamics, dividing the modeling techniques into two complimentary classes: those with artificial dynamics and those with biophysically realistic dynamics (Fig. 8.4b). As we will see, models from each of these two classes are able to reproduce important aspects of neural activity and system function that have been observed in a range of physiological and behavioral experiments.

8.3.2.1 *Artificial models*

One of the earliest mathematical models of neural activity whatsoever was proposed in the mid-1940s by Warren McCulloch and Walter Pitts to describe the logical functioning of an individual neuron (*435*). Known as the MP neuron, their model accepted binary inputs, combined these inputs using linear weights, and produced a binary output reflecting whether or not the weighted sum of inputs exceeded a given threshold (Fig. 8.4b). Albeit a simple caricature of neuronal dynamics, this model has been shown to reproduce some important qualitative features of neuronal activity, including the linear summation of excitatory inputs (*125*) and the "all-or-none" response to the resulting integrated signal (*214*). Moreover, by connecting the inputs and outputs of multiple MP

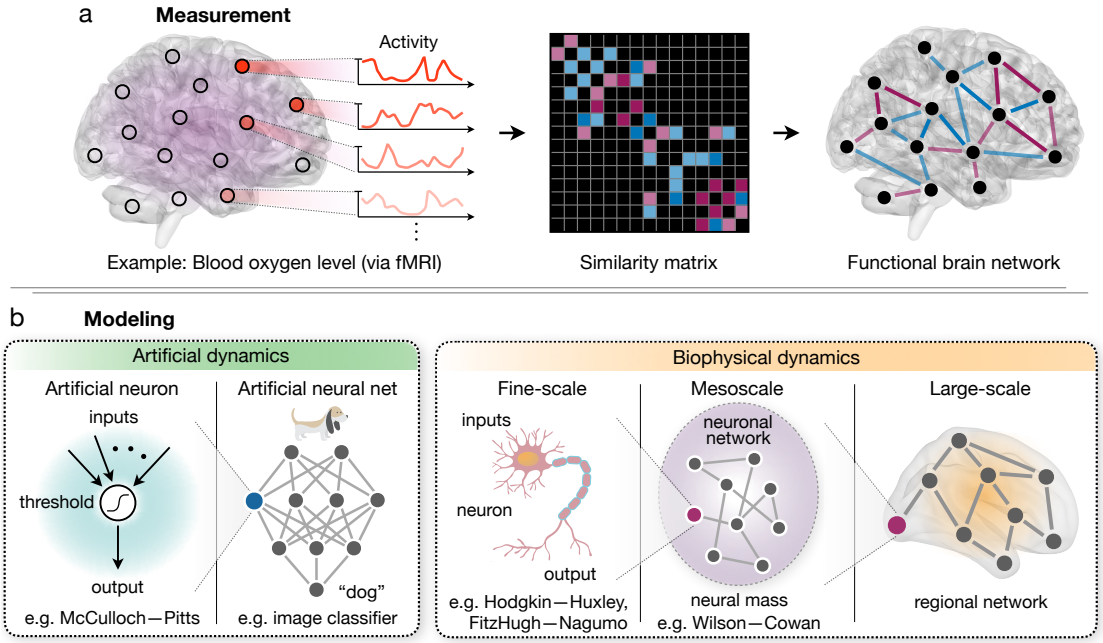


Figure 8.4: Measuring and modeling brain network function. (*a*) The measurement of brain network function begins with experimental data specifying the activity of neurons or brain regions. As an example, we consider variations in blood oxygen level in different parts of the brain measured via fMRI. Calculating the similarity (e.g., correlation or synchronization) between pairs of activity time series, one arrives at a similarity matrix. This matrix, in turn, defines a functional brain network constructed from our original measurements of neural activity. (*b*) We divide models of neural activity into two classes: abstract models with artificial dynamics (*left*) and biophysical models with realistic dynamics (*right*). Models of artificial neurons, such as the MP neuron, typically take in a weighted combination of inputs and pass the inputs through a nonlinear threshold function to generate an output. Networks of artificial neurons, from deep neural networks to Hopfield networks, have been shown to reproduce key aspects of human information processing, such as learning from examples and storing memories. By contrast, biophysical models of individual neurons, such as the Hodgkin–Huxley or FitHugh–Nagumo models, capture realistic functional features such as the propagation of the nerve impulse. When interconnected with artificial synapses, researchers are able to simulate entire neuronal networks. Complimentary mesoscale approaches, including neural mass models such as the Wilson–Cowan model, average over all neurons in a population to derive a mean firing rate. To simulate the large-scale activity of an entire brain, researchers use neural mass models to represent brain regions and embed them into a network with connectivity derived from measurements of neural tracts (e.g., as measured via DTI).

neurons, researchers have achieved deep insights about how brain networks perform basic cognitive functions. For example, soon after the introduction of the MP model, researchers demonstrated that networks of artificial neurons could be used to represent any Boolean function (i.e., any function mapping a list of binary variables to a binary output), thereby establishing the basic capability of neural networks to perform logical computations (*596*).

While their ability to perform basic computations was quickly realized, it was not clear at the outset whether artificial neural networks could reproduce other cognitive functions, such as the ability to learn or store memories. The former was established by Frank Rosenblatt in 1957, when he showed that the weights on the inputs to an MP neuron could be tuned such that the output defines a binary classifier. Known as the perceptron, this algorithm enabled a single MP neuron to segregate incoming data into one of two classes by learning from past examples. This remarkable result directly inspired more advanced learning algorithms, including support vector machines (*300*) and artificial neural networks (*374*), effectively setting in motion the study of machine learning. Today, deep neural networks, consisting of multiple layers of artificial neurons feeding in one direction from the input layer to the output layer (Fig. 8.4b), are able to learn a wide range of impressive cognitive functions that we have come to expect from the brain (*588*). While the list of applications is ever-expanding, deep neural networks have been used to process and identify images of objects, scenes, and people (*200*); recognize, interpret, and respond to spoken language (*315*); and formulate strategies and make decisions in adversarial settings (*614*).

In addition to performing computations and learning from examples, the physicist John Hopfield showed in 1982 that neural networks can also store and recall memories. Specifically, Hopfield demonstrated that the synaptic weights connecting a set of MP neurons could be adjusted in a Hebbian fashion such that the network is able to "memorize" a number of desired activity states (*322*) (i.e., configurations of the network in which each neuron is either active or inactive). Notably, the number of memorized states grows linearly with the number of neurons in the network (*473*), and errors in recall often yield states that are semantically similar to the target state, a phenomenon commonly observed in humans (*308*). Interestingly, the memorized activity states can be interpreted as local minima of an associated energy function, making each Hopfield network equivalent to an Ising model at zero temperature (*113*). More recently, Ising-like models have also been used to explain the critical or avalanche-like behavior of activity in neural ensembles (*456*), which is thought to support adaptation to environmental changes (*713*), information storage (*291*), optimal information transmission (*74*), maximal dynamic range (*369*, *608*), and computational power (*78*). Further building upon this connection to statistical mechanics, scientists have recently used maximum entropy techniques to construct data-based models of neuronal dynamics. These maximum entropy models, which are equivalent to networks of Ising spins with specially-chosen external fields and interaction strengths, have been shown to predict the observed long-range correlations within naturally occurring networks of neurons and brain regions (*241*, *589*). Together, artificial models of neural dynamics, from simple MP neurons to artificial neural networks and data-driven maximum entropy models, continue to inform our understanding of brain networks as information processing systems.

8.3.2.2 *Biophysical models*

While artificial models continue to generate insights about the nature of neural computation, they only vaguely resemble the complex biophysical mechanisms that guide observable neural activity. Among the first biophysically realistic models of the electrical behavior of an individual neuron was achieved nearly a decade after the introduction of the MP neuron by physiologists Alan Lloyd Hodgkin and Andrew Fielding Huxley (*316*). Beginning from a principled description of the initiation and propagation of action potentials in living neurons, the Hodgkin–Huxley (HH) model explains important qualitative aspects of neuronal behavior (*596*), including the spontaneous emergence of limit cycles or oscillations in activity (*393*) and the presence of a Hopf bifurcation in the neuronal firing rate, which is thought to underlie the all-or-none principle (*316*) (Fig. 8.4c). Subsequent extensions of the HH model expand biophysical realism by incorporating multiple ion channel populations (*311*), the complex geometries of dendrites and axons (*528*), and more realistic stochastic dynamics yielding thermodynamic and hybrid HH models (*20*, *497*). Concurrent with these descriptive improvements, several simplified neuronal models were also developed, including the notable FitzHugh–Nagumo model (*218*, *466*), facilitating efficient large-scale simulations of groups of neurons.

Simplifications in neuronal modeling, paired with fine-scale measurements of the synaptic wiring in several animals, have spurred large-scale simulations of real neuronal circuits (Fig. 8.4b). For example, on the heels of mapping the entire *C. elegans* connectome (*705*), researchers began simulating the 302-neuron network at the cellular level (*482*), eventually even including the nematode's entire muscular system and representations of its physical environment (*114*). Despite these and other efforts simulating the *Drosophila* brain (*24*) and the rat's neocortical column (*427*), it remains unclear how networks of neurons combine to generate the complex range of behaviors observed even in these relatively simple organisms. This contrast between the simplicity of neuronal dynamics and the apparent complexity of large-scale neural behavior hints at the crucial role of emergence. To understand how macroscopic behaviors emerge within groups of neurons, researchers began developing mean-field descriptions of large neuronal populations. Known as neural mass models, these efforts culminated in the foundational Wilson–Cowan (WC) model of population dynamics (*710*). Whereas previous neural mass models only considered excitatory interactions between neurons, Wilson and Cowan also included inhibitory interactions, thereby enabling the WS model to predict the collective neural oscillations observed in experiments as well as the emergence of other key properties of neural behavior, including the existence of multiple stable states and hysteresis in the neural response to stimuli (*710*). This progress was further extended to include spatial fluctuations in activity, yielding neural field models that exhibit other behaviors typically observed in the brain, including regions of localized activity (*371*) and traveling waves (*527*).

In much the same way that neuronal circuits have been modeled using observable synaptic wiring in animals, one could imagine simulating a network of neural mass

models whose connections are drawn based on non-invasive measures of regional connectivity in humans. By doing so, researchers are now able to simulate whole sections of the human brain (Fig. 8.4c), opening the door for comparisons with experimental measurements of regional activity. Precisely this approach has driven a deeper understanding of the structure-function relationship, including the demonstration that the broad spectrum of MEG/EEG recordings of electrical activity can be reproduced by networked models of neural masses (*168*) and that the functional connectivity within such recordings depends critically on the coupling strength between neural masses (*169*). To facilitate large-scale simulations of the entire human brain, researchers have frequently turned to the Kuramoto model of oscillatory dynamics as a simplified neural mass model (*384*, *385*). These efforts have provided insights about the spontaneous synchronization of neural oscillations (*698*), a phenomenon which is thought to play a critical role in neural communication (*229*), information processing (*500*), and motor coordination (*590*). Moreover, by embedding Kuramoto oscillators into a realistic map of the human connectome, researchers have shown that even this simple model is able to reproduce the patterned fluctuations in activity and long-range correlations observed in fMRI data (*119*). Detailed biophysical models of neural dynamics, from descriptions of the electrical activity of individual neurons to networked neural mass models simulating the entire brain, continue to inform our understanding of how collective neural behavior and high-level cognitive functions arise from the brain's underlying physical circuitry.

8.3.3 *The future of brain network function*

Over the last two centuries, our understanding of the brain's functional organization and information processing capabilities has progressed immensely. Despite this progress, the modern neuroscientist remains fundamentally limited by the experimental and theoretical tools at their disposal (*521*, *522*). Invasive techniques such as intracranial electrocorticography, and even minimally invasive techniques such as stereotactic electroencephalography (sEEG) (**Ch8-todaro2018mapping**, *46*, *138*), provide immense precision in mapping human brain dynamics, but remain constrained to patients with medically refractory epilepsy. Other noninvasive imaging techniques all suffer from trade-offs between spatial and temporal resolution (*443*); methods that directly measure electromagnetic signals (e.g., EEG and MEG) have high temporal resolution but low spatial resolution, while measurements of blood flow and metabolic activity (e.g., via fMRI or PET) have relatively high spatial accuracy but poor resolution in time. Even fMRI – widely considered the standard for high spatial resolution in humans – integrates signals over hundreds of thousands of neurons and several seconds (*9*). Consequently, any changes in neural activity that occur over tens of thousands of neurons or even over the span of a second are imperceivable on a standard fMRI scan.

To improve the precision of functional neuroimaging (fMRI in particular), recent efforts have leveraged modern advances in image processing to strengthen the signal and reduce background noise. For example, to minimize the inevitable effects of

head movements and fluctuations in blood flow during scanning, fMRI signals are increasingly corrected using techniques similar to image stabilization in video cameras (*146*). Additionally, in order to draw general conclusions from neuroimaging results across a group of subjects, impressive strides have been made to correct for inter-subject heterogeneities in brain structure (*35*). Together, advances in image processing have begun to push neuroimaging from a tool exclusively used for academic research to one that can aid in the diagnosis and treatment of psychiatric disorders such as schizophrenia and Alzheimer's disease.

Beyond data collection, data analysis and models in network neuroscience have historically been limited to dyadic relationships between neural elements, such as synapses connecting pairs of neurons or Pearson correlations between pairs of brain regions (*60*, *63*). While these dyadic notions of connectivity have provided important insights about the brain's circuitry, mounting evidence suggests that higher-order interactions between three or more elements are also crucial for understanding the large-scale behavior of entire brain networks (*16*, *241*, *616*). In order to study these higher-order connections, recent efforts have focused on generalizing traditional definitions and intuitions from network science, primarily by adopting methods from algebraic topology (*252*). One notable approach, known as persistent homology, has allowed researchers to extrapolate conclusions about neural activity across scales, escape the problem of selecting appropriate thresholds for functional edge strengths (*251*), and extract principled mesoscale features of network organization (*552*, *616*).

Efforts have also been made to expand traditional metrics of functional connectivity, which are typically based on correlation, to include more sophisticated notions of causality (*79*). Since causality reflects the flow of information in a network from one element to another, efforts which aim to uncover causal relationships between neurons and brain regions have naturally drawn inspiration from concepts in information theory (see below) (*66*). From mutual information to transfer entropy, information theoretic notions of functional connectivity are increasingly being used to quantify the flow of information in the brain (*370*, *500*, *731*). These measures of causality, in turn, have real-world implications for controlling brain networks and intervening to treat neurological disease and psychiatric disorders.

Information theory and network neuroscience. At its core the brain is an information processing system, having evolved over millions of years to encode and manipulate a continuous stream of sensory signals (*560*). As such, information theory – the science of how signals are encoded and processed – provides a compelling lens through which to study the brain's function (*161*). Information theory began with the 1948 paper "A Mathematical Theory of Communication," wherein Claude Shannon proposed the entropy of a signal as the natural measure of its information content and derived fundamental limits on the information capacity of a communication channel (*603*). Soon after, MacKay and McCulloch adapted the concept of channel capacity to obtain limits on the rate at which one neuron

can transmit information to another (*421*), sparking the study of information flow in the brain. Subsequent work by Attneave and Barlow proposed the idea that neural activity is optimized for the transmission of sensory information (*32*, *53*), providing the foundation for future investigations of neural coding (*1*, *560*).

Despite these initial efforts bridging information theory and neuroscience, progress slowed primarily due to difficulties obtaining unbiased information estimates from neural systems. Improvements in experimental techniques, however, eventually sparked renewed interest (*682*), spurring the introduction of robust methods for estimating information theoretic quantities (*470*, *502*, *640*). On the basis of these advancements, information theory has once again become a powerful tool for the network neuroscientist. Recent attempts, for instance, to uncover causal relationships between neural elements have successfully adapted notions of information flow, such as mutual information and transfer entropy (*593*, *685*). At the same time, efforts to understand large-scale correlations within neuronal populations have utilized the principle of maximum entropy (*339*), resulting in Ising-like models of collective neural behavior (*241*, *589*). As information theory becomes increasingly integrated into the fabric of neuroscience, physicists are uniquely positioned to pioneer exciting new techniques for investigating the nature of information processing in the brain.

8.4 PERTURBATIONS AND THE PHYSICS OF BRAIN NETWORK CONTROL

Thus far, we have examined what is known about the structural circuitry connecting neural components in the brain as well as the dynamical laws governing the interactions between these components. An ultimate test of our understanding, however, lies in our ability to intervene and shift the brain's dynamics to facilitate desirable behaviors. An important implication of the brain's networked structure is that localized perturbations (e.g., targeted lesions or stimulation) do not just yield localized effects – they also induce indirect effects that propagate along neural pathways (*412*, *436*). In this way, the task of controlling brain dynamics requires knowledge of how signals transmit along the brain's structural wires, making the problem inherently one of network control (*409*). Building upon targeted lesioning experiments in animals and clinical interventions in humans, efforts toward a theory of network control in the brain have recently taken shape, inspiring several fundamental questions (*586*). Are brain networks designed to facilitate control (*367*)? What are the principles that allow brain networks to control themselves toward desired activity states (*284*, *341*)? Can we leverage these principles to inform stimulation-based therapies for neurological diseases and psychiatric disorders (*319*, *440*, *464*, *655*)? To address these questions, here we review the current frontiers in the physics of brain network control.

8.4.1 *Targeted perturbations and clinical interventions*

The first attempts to systematically control brain dynamics date to the early 19th century, when Marie-Jean-Pierre Flourens noticed that targeted lesions to the brain in living rabbits and pigeons yielded specific changes in the animals' perception, motor coordination, and behavior (*220*). These efforts, in conjunction with other targeted lesioning experiments in animals (*108*, *503*), supported the notion of functional localization – the theory that specific cognitive functions are supported by specific parts of the brain. In humans, evidence for functional localization has typically relied on patients with localized brain damage (e.g., due to a stroke or head trauma). Historical studies of this kind have revealed, for instance, that damage to one half of the occipital lobe often induces blindness in the opposite field of vision (*318*) and that lesions in the frontal lobe can result in memory loss and an increase in impulsivity and risk taking (*495*). More recently, advances in non-invasive stimulation techniques such as transcranial magnetic stimulation (TMS) (*694*), which induces "transient" lesions by disrupting the brain's normal electrical activity, have opened the door for the control of localized brain functions, including perception (*17*), learning (*511*), language processing (*510*), and attention (*695*). These non-invasive transcranial techniques have been supplemented by more invasive deep brain stimulation (DBS) methods to provide targeted therapies for a number of psychiatric and neurological disorders (*382*, *694*). By focusing electromagnetic stimulation on the brain regions associated with specific disorders, both TMS and DBS have been used to treat Parkinson's disease, epilepsy, depression, and schizophrenia, among other disorders that are resistant to traditional therapies (*243*, *518*) (Fig. 8.5a). Despite these therapeutic benefits, it remains unclear exactly how and why TMS and DBS are so effective (*382*, *436*); however, recent evidence suggests that the answers may rely on a deeper understanding of the indirect effects of stimulation that are mediated by the brain's physical circuitry (*579*, *654*).

With the recent development of whole-brain neuroimaging methods such as fMRI, evidence continues to mount that brain regions are heavily interdependent on one another, often working in unison to process information and formulate responses (*179*, *677*). In a particularly clear demonstration of the brain's functional integration, Anthony Randall McIntosh and colleagues trained human subjects to associate an auditory stimulus with a visual event. Later, when the auditory stimulus was presented alone, the investigators observed increased activity in the occipital lobe, more traditionally thought of as being reserved for visual processing (*725*). Experiments such as these reveal how activity or stimulation in one part of the brain can propagate along neural pathways to induce activity in other distant parts. To understand the system-wide impacts of targeted stimulation, researchers have increasingly drawn upon network models of brain dynamics (*579*, *654*). These efforts have resulted in the identification of neural circuits, rather than isolated regions, that are critical for reducing the symptoms of Parkinson's disease (*142*, *579*). Similar network-based approaches are also being used to suppress epileptic seizures using DBS (*77*), non-invasively treat depression using TMS (*363*), and modulate consciousness during surgery using anesthesia (*143*). Moreover,

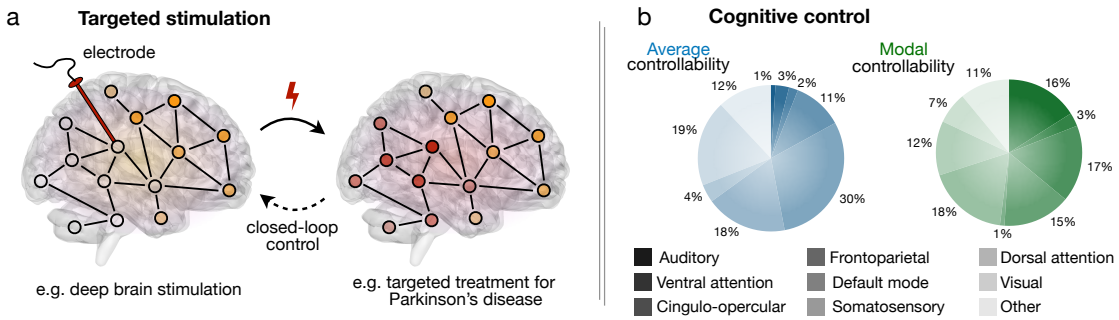


Figure 8.5: Targeted perturbations and brain network control. (*a*) Methods for targeted control are used in the study, design, and optimization of external control processes, such as transcranial magnetic stimulation and deep brain stimulation. These targeted perturbations of neural activity are being utilized in clinical settings to treat major depression, epilepsy, and Parkinson's disease. By simultaneously stimulating and measuring neural activity, researchers can now perform closed-loop control, continuously updating stimulation strategies in real time. (*b*) Controllability metrics provide summary statistics regarding the ease with which a given node can enact influence on the network. Two common metrics are the average controllability, which assesses the ease of moving the system to all nearby states, and the modal controllability, which assesses the ability to move the system to distant states (see Fig. 8.6). Notions of controllability have proven useful in the study of the brain's internal control processes, such as homeostatic regulation and cognitive control. For example, the human brain displays marked levels of both average and modal controllability, and the proportion of average and modal controllers differs across cognitive systems, suggesting the capacity for a diverse repertoire of dynamics (*284*).

by stimulating and recording neural activity in several brain regions simultaneously, researchers have achieved closed-loop strategies for dynamically updating targeted treatments (*302*, *320*) (Fig. 8.5a). Meanwhile, clinical applications are increasingly being informed by detailed computational simulations of perturbations to specific brain regions, typically employing networked biophysical models such as those discussed in the previous section (*164*, *578*). Together, these real-world and computational studies of targeted stimulation have opened the door for sophisticated strategies that aim to shift neural activity with the ultimate goal of guiding healthy cognitive function.

8.4.2 *Network control in the brain*

To inform strategies for targeted stimulation and brain network control, it helps to draw upon existing tools from control theory in mathematics and intuitions from cognitive control in psychology. Given a mathematical model of a system, control theory seeks to understand how the system can be influenced such that it moves toward a desired state (*335*, *409*) (see Fig. 8.6). Cognitive control, on the other hand, encompasses a broad class of processes by which the brain enacts control over itself, typically to achieve an abstract goal or desired response (*538*). For example, dating to the early 1970s neurophysiological studies revealed that the act of holding an object in working memory induces a sustained neural response in the prefrontal cortex (*235*, *259*). In fact, the prefrontal cortex is now believed to play a key role in many cognitive

control processes, from the representation of complex goal-directed behaviors (*69*) to the support of flexible responses to changes in the environment (*186*). But how do these notions of cognitive control (as defined by psychologists and cognitive neuroscientists) compare to theories of network control (as defined by physicists and engineers)? Furthermore, how can knowledge of the brain's intrinsic control processes inform targeted therapies for mental illness?

To address these questions, we begin by comparing cognitive notions of intrinsic control with theoretical measures of control and controllability in brain networks (see Fig. 8.6). It is interesting, for example, to ask which brain regions are most capable of inducing desired neural responses in other brain regions that are responsible for common functions such as vision, audition, and motor coordination. Toward this end, Gu *et al.* used methods from control theory to demonstrate that the strongest driver nodes corresponded to brain regions with high communicability – or many topological paths through the brain network – to the target brain regions (*285*). In a related study, Betzel *et al.* used the structural wiring of the brain to simulate transitions between commonly observed activity states (*85*). They found that optimal control nodes tended to have high degree in the network, and that when this rich-club of hub regions was destroyed by simulated lesioning, the ability of the brain to make common transitions was significantly reduced.

In addition to studying the roles of specific control trajectories, complementary approaches have considered trajectory-independent metrics such as the average and modal controllabilities discussed in Fig. 8.6 (*512*). By comparing control theoretic measures of node controllability with the cognitive functions associated with each brain region, researchers have observed that different types of controllers are located in distinct areas of the brain (Fig. 8.5b) (*284*). For example, brain regions with strong average controllability are disproportionately located in the default mode system, which is associated with baseline neural activity; meanwhile, strong modal controllers are primarily located in cognitive control systems. These observations are particularly interesting because they suggest that regions associated with the default mode are optimally positioned to push the system into many easily reachable states, while regions associated with cognitive control are optimally positioned to steer the system toward distant states.

As a final layer of abstraction, rather than studying the controllabilities of specific brain regions, one could envision averaging over all regions to quantify the mean controllability of an entire brain network. Interestingly, by taking precisely this approach, Tang *et al.* established that brain networks as a whole are finely tuned to maximize both average and modal controllability, thereby supporting a diverse range of possible control strategies (*653*). Furthermore, by comparing subjects in different stages of adolescence, the researchers found that brain network controllability increases with age, suggesting that neural circuitry evolves over time to support increasingly complex dynamics. In related studies, metrics of network controllability were found to differ by sex (**Ch8-cornblath2018sex**) and to be altered in individuals with high genetic risk for bipolar disorder (*341*). Taken together, these results demonstrate that network measures

Linear control and network controllability. To investigate the principles of control in the brain, it is useful to understand the theory of network control generally. In network control, the system in question typically comprises a complex web of interacting components, and the goal is to drive this networked system toward a desired state by influencing a select number of input nodes (*409*). The starting point for most control theoretic problems is the linear time-invariant control system $\mathbf{x}(t+1) = \mathbf{A}\mathbf{x}(t) + \mathbf{u}(t)$, where $\mathbf{x}(t)$ defines the state of the system (e.g., the BOLD signal measured by fMRI), $\mathbf{A}$ is the interaction matrix (e.g., white matter tracts estimated using DTI), and $\mathbf{u}(t)$ defines the input signal (e.g., electromagnetic stimulation using TMS or DBS) (*352*). Such a system is said to be *controllable* if it can be driven to any desired state. Often, however, many naturally occurring networks that are theoretically controllable cannot be steered to certain states due to limitations on control resources (*376*, *410*), motivating the introduction of control strategies $\mathbf{u}^*(t)$ that minimize the so-called *control energy* $E(\mathbf{u}) = \sum_{t=0}^{\infty} |\mathbf{u}(t)|_2^2$.

By limiting the control input to a single node, we can quantify the ability of that node to steer the dynamics of the entire system. For example, the *average controllability* of a node represents its capacity to drive the network to many nearby states (*284*), while a node's *modal controllability* quantifies its ability to push the network toward distant hard-to-reach states (*512*) (see figure). Averaging these metrics over all nodes in a system, one can estimate the inherent controllability of an entire network itself. Control theoretic efforts such as these have only recently been applied to understand the locomotion of the nematode (*715*) and the networked behavior of the brain more broadly (*412*, *586*, *652*), promising new strategies for stimulation-based therapies and fresh insights about the brain's capacity for intrinsic control.

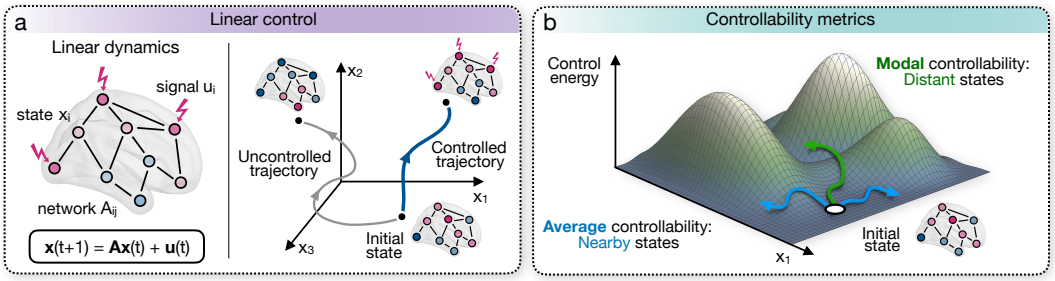


Figure 8.6: Control theory in the brain. (*a*) Linear control theory describes how to influence a linear system to move along a desired trajectory. (*b*) Controllability metrics, including average and modal controllability, quantify the ease with which a given system can be controlled.

of optimal control and controllability correspond closely to existing notions of intrinsic and cognitive control in neuroscience. This close correspondence, in turn, suggests that network control theory, by taking into account the complex wiring of the brain, has the promise to enrich our understanding of the brain's control principles (*652*).

8.4.3 *The future of brain network control*

Throughout this section, we have focused primarily on targeted therapies that rely on the coarse-grained stimulation of entire brain regions and simple control strategies that assume idealized linear dynamics. Emerging efforts in neuroscience and control theory, however, are opening the door for a number of significant improvements, including: (i) techniques for fine-scale control of neural activity (*4*, *181*, *282*, *287*), even down to the level of individual neurons (*540*, *558*), (ii) systems identification approaches that allow for the incorporation of effective connectivity measurements to inform control, superseding solely structural explanations (*72*), and (iii) generalizations of linear control theory that include more realistic nonlinear dynamics (*158*, *377*). Among recent advances in the manipulation of fine-scale neural activity, arguably the most promising tool is optogenetics, which offers millisecond-scale optical control of specific cell types within the brains of conscious animals (*4*, *181*). Its striking precision (*287*), in some cases even down to single-cell resolution (*540*, *558*), has enabled researchers to investigate the nature of causal signals between neurons and to study how these signals give rise to qualitative changes in animal behavior (*282*).

While linear control theory continues to provide critical insights about how signals propagate along the brain's structural wiring (*85*, *284*, *285*, *367*), interactions between neural components, from individual neurons to entire brain regions, are highly nonlinear (Fig. 8.4b) (*106*). Initial efforts to develop a theory of nonlinear control, dating as early as the 1970s (*298*, *306*, *644*), quickly converged on the conclusion that results as strong and general as those derived for linear dynamics could not be obtained for a general nonlinear system (*409*). Fortunately, concerted theoretic efforts have led to weaker notions of nonlinear controllability (*157*), notable among which are techniques for linearizing nonlinear systems around stable equilibrium states (*158*, *377*) and methods for leveraging the symmetries of a system (*703*) such as repeated network motifs to simplify control strategies (*333*). Additional efforts have utilized advances in computing power to simulate the effects of external perturbations across a range of model systems, including networks of FitzHugh–Nagumo neurons (*703*), Wilson–Cowan neural masses (*464*), and Kuramoto oscillators (*145*) as well as artificial neural networks such as the Ising model (*416*, *417*). Together, recent advances in high-precision neural stimulation like optogenetics and our emerging understanding of the principles governing nonlinear control are pushing the boundaries of what is considered possible in the investigation of neural activity. Targeted control of the brain's complex behavior – once considered a topic of science fiction – now has the promise to shape targeted therapies for a range of psychiatric and neurological disorders.

8.5 conclusions and future directions in the neurophysics of brain networks

The intricate inner workings of the brain remains one of the greatest mysteries defying resolution by contemporary scientific inquiry. On the heels of decades of effort investigating the functions of the brain's individual components (*19*), from neurons to neuronal ensembles and large-scale brain regions, conclusive evidence points to the need for maps and models of the interactions between these components in order to fundamentally understand the brain's ensemble dynamics, circuit function, and emergent behavior (*63*, *151*). Here we reviewed recent advances toward meeting this challenge with an eclectic array of curios from the physicist's cabinet: statistical mechanics of complex networks, thermodynamics, information theory, dynamical systems theory, and control theory. In the course of our exposition, we considered the principles of small-worldness (*57*), interconnected high-degree hubs (*679*), modularity (*624*), and spatial embedding (*637*) that provide useful explanations for the architecture of structural brain networks. We then saw these same principles reflected in the organization of long-range functional connectivity supporting information dissemination, and the computations that can result therefrom (*36*, *500*). As with any physical system, a natural next step is to probe the validity of our descriptive and explanatory models using perturbative approaches both in theory and experiment. Thus, we next summarized the utility of network control theory in offering insights into internal control processes such as homeostatic regulation and cognitive control, as well as external control processes such as neurostimulation, which are currently being used to treat multiple disorders of mental health (*652*).

Throughout the exposition, we described current frontiers in the investigation of brain network structure, function, and control. Although we will not reiterate those points here, we do wish to offer the sentiment that, while the empirical advances laying the foundation of the field have spanned several decades, the network physics of the brain is an incredibly young area, rich with opportunities for discovery. And perhaps – with a bit of courage – we may even begin to provide an empirical constitution to the deeper philosophical questions that humans have wrestled with for millennia: What makes us unique and different from non-human animals (*367*, *680*)? How do we represent abstract concepts such as value to ourselves (*519*) and others (*191*)? How are representations transmitted throughout the brain or reconfigured based on new knowledge (*155*)? What makes a mind from a brain? Physicists, the brain is calling you.

9

NON-EQUILIBRIUM DYNAMICS AND ENTROPY PRODUCTION IN THE HUMAN BRAIN

This chapter contains work from Lynn, Christopher W., Eli J. Cornblath, Lia Papadopoulos, Maxwell A. Bertolero, and Danielle S. Bassett. "Non-equilibrium dynamics and entropy production in the human brain." In preparation.

Abstract

Living systems fundamentally exist out of equilibrium (*594*), producing entropy in the environment by maintaining order and performing biological functions. However, whereas non-equilibrium processes are critical for molecular and cellular operations (*103*, *197*, *328*, *389*, *441*, *641*, *718*), it remains unclear if and how non-equilibrium dynamics manifest at macroscopic scales. Here we present a framework to probe for non-equilibrium dynamics and quantify entropy production in complex living systems. Applying our method to whole-brain imaging data, we demonstrate that the human brain fundamentally functions out of equilibrium. Moreover, the brain produces more entropy – operating further from equilibrium – during periods of physical and cognitive exertion. Comparing against simulated dynamics, we show that this capacity of the brain to operate at different distances from equilibrium resembles tuning the temperature of an asymmetric Ising model. Together, these results provide a general tool for probing and quantifying non-equilibrium dynamics at macroscopic scales.

9.1 INTRODUCTION

The functions that support life – from processing information to generating forces and maintaining order – require organisms to operate far from thermodynamic equilibrium (*255*, *594*). For a system at equilibrium, the fluxes of transitions between different states vanish (Fig. 9.1a), a property known as detailed balance; the system ceases to produce entropy and its dynamics become reversible in time. By contrast, living systems exhibit net fluxes between states or configurations (Fig. 9.1b), thereby breaking detailed balance and establishing an arrow of time (*255*). Critically, such non-equilibrium dynamics lead to the production of entropy, a fact first recognized by Sadi Carnot in his pioneering studies of irreversible processes (*124*). At the molecular scale, enzymatic activity drives non-equilibrium processes that are crucial for intracellular transport (*103*), high-fidelity transcription (*718*), and biochemical patterning (*328*). At the level of cells and subcellular structures, non-equilibrium activity enables sensing (*441*), adaptation

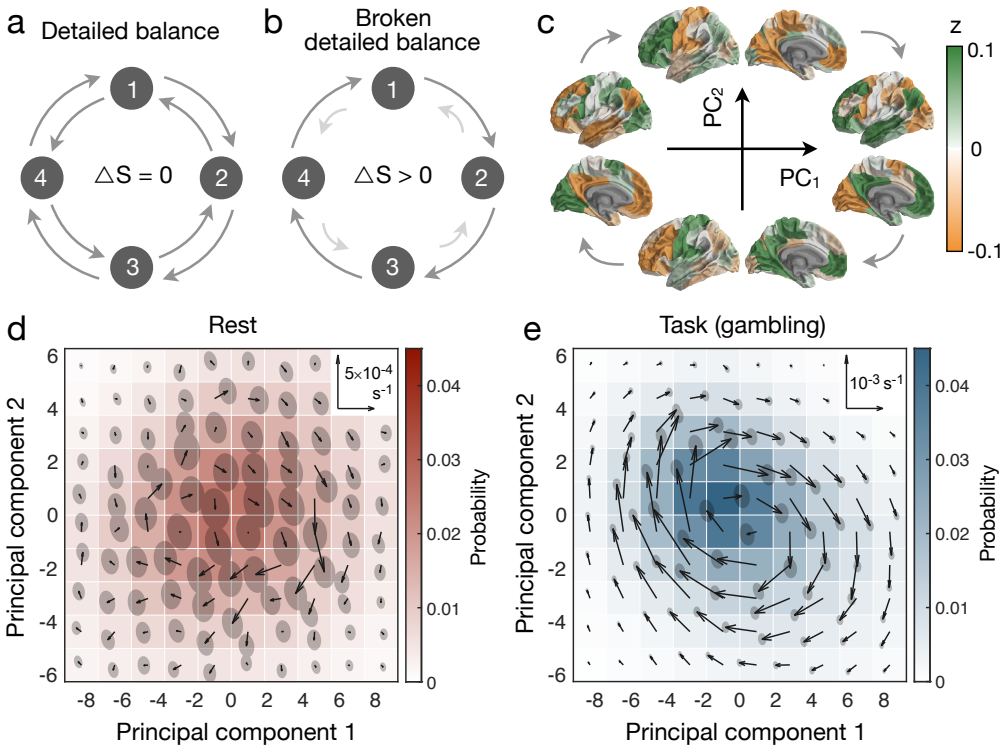


Figure 9.1: Macroscopic non-equilibrium dynamics in the brain. (*a-b*) A simple four-state system, with states represented as circles and transition rates as arrows. (*a*) At equilibrium, there are no net fluxes of transitions between states – a condition known as detailed balance – and the system does not produce entropy. (*b*) Systems that are out of equilibrium exhibit net fluxes of transitions between states, breaking detailed balance and producing entropy in the environment. (*c*) Brain states defined by the first two principal components of the neuroimaging time-series, calculated for all time points and all subjects. Colors indicate the z-scored activation of different brain regions, ranging from high-amplitude activity (green) to low-amplitude activity (orange). Arrows represent possible fluxes between states. (*d-e*) Probability distribution (color) and net fluxes between states (arrows) for neural dynamics at rest (*d*) and during a gambling task (*e*). In order to use the same axes in panels **d** and **e**, the dynamics are projected onto the first two principal components of the combined rest and gambling time-series data. The flux scale is indicated in the upper right, and the disks represent two-standard-deviation confidence intervals for fluxes estimated using trajectory bootstrapping (*604*) (see Methods; Fig. 9.5).

(*389*), force generation (*197*), and structural organization (*641*). However, despite the importance of non-equilibrium processes at small scales, there remain fundamental questions concerning how – and even whether – non-equilibrium dynamics unfold in macroscopic systems composed of many interacting components. Indeed, the amount of entropy produced by a system can only decrease with coarse-graining (*208*), leading to the possibility that complex living systems, despite operating far from equilibrium at small scales, may appear to regain equilibrium at large scales (*201*).

Here we study the non-equilibrium nature of the human brain, a complex web of interacting neurons that, despite accounting for only 2% of the body's weight,

consumes up to 20% of its energy (*295*). This immense metabolic expenditure drives an array of non-equilibrium processes at the microscale, including molecular cycles (*486*), cellular housekeeping (*194*), and neuronal firing (*206*). But does the brain function out of equilibrium at large scales? And if so, does the non-equilibrium nature of the brain vary with physical or cognitive demands? To answer these questions, we develop a framework to probe for non-equilibrium dynamics and quantify entropy production in complex living systems.

9.2 FLUXES AND BROKEN DETAILED BALANCE IN THE BRAIN

We analyze whole-brain dynamics from 590 healthy adults both at rest and across a suite of seven cognitive tasks, recorded using functional magnetic resonance imaging (fMRI) as part of the Human Connectome Project (*678*). The time-series data consist of blood-oxygen-level-dependent (BOLD) fMRI signals from 100 cortical parcels (*662*) (see Methods), which we concatenate across all subjects. To visualize the neural dynamics, we project the time series onto the first two principal components, which are calculated for all data points and all subjects (Fig. 9.1c). In fact, this projection defines a natural low-dimensional state space (*165*), capturing over 30% of the variance in the neural activity (Fig. 9.6). One can then probe for non-equilibrium dynamics by calculating the net fluxes of transitions between different regions of state space (*67*) (see Methods). During rest scans, wherein subjects are instructed to remain still with their eyes open, we find that the brain exhibits net fluxes between states (Fig. 9.1d), thereby breaking detailed balance and departing from equilibrium. However, these resting-state fluxes are randomly oriented and weak compared to statistical errors. For comparison, we consider task scans, wherein subjects respond to stimuli and commands that require attention, cognitive effort, and computations. For example, during a gambling task in which subjects play a card guessing game for monetary reward, the brain's dynamics form a distinct loop of fluxes (Fig. 9.1e) that are nearly an order of magnitude stronger than those present during rest. Such closed loops of flux are a characteristic feature of non-equilibrium steady-state systems (*729*), and we verify that the brain does operate in a stochastic steady state (Fig. 9.7). Finally, we confirm that if the neural dynamics are shuffled in time – thereby destroying the temporal order of the system – then the fluxes between states vanish and equilibrium is restored (Fig. 9.8). Together, these results demonstrate that the brain fundamentally operates out of equilibrium at large scales, and moreover, that the nature of this non-equilibrium behavior depends critically on the cognitive function being performed.

9.3 NON-EQUILIBRIUM DYNAMICS IN AN ASYMMETRIC ISING MODEL

To understand how non-equilibrium dynamics arise at large scales, it is helpful to consider a canonical model of stochastic dynamics in complex systems. In the Ising model, the interactions between spins are typically constrained to be symmetric, yield-

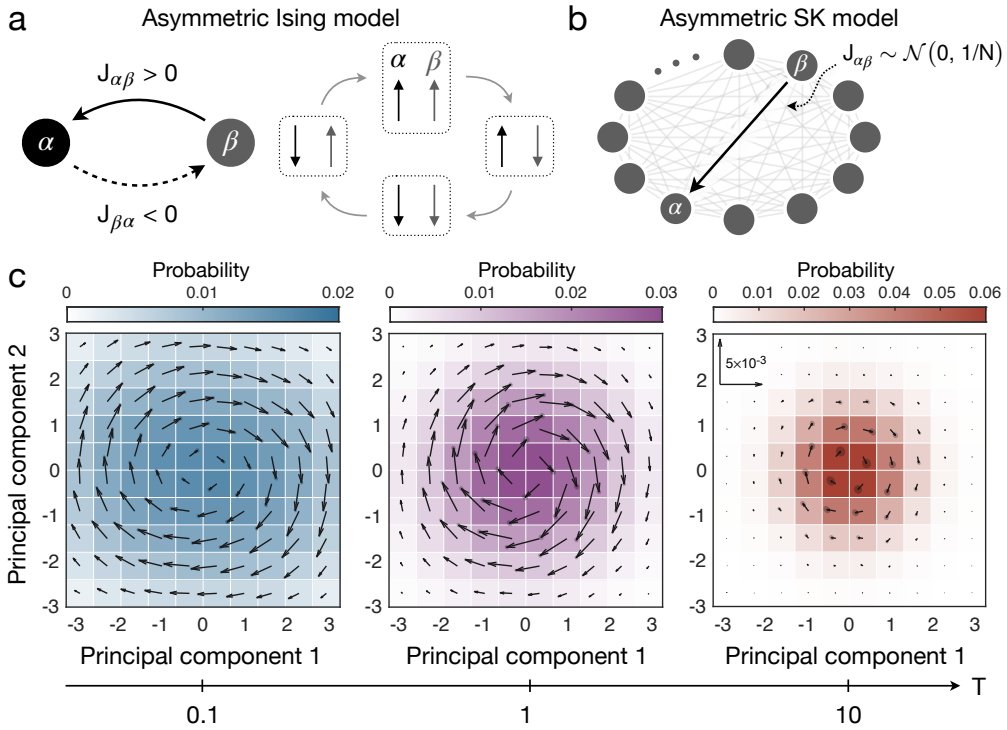


Figure 9.2: Simulating complex non-equilibrium dynamics using an asymmetric Ising model. (*a*) Two-spin Ising model with asymmetric interactions (left), where the interaction $J_{\alpha\beta}$ represents the strength of the influence of spin β on spin α. Simulating the model with synchronous updates, the system exhibits a clear loop of flux between configurations (right). (*b*) Asymmetric version of the Sherrington-Kirkpatrick (SK) model, wherein directed interactions are drawn independently from a zero-mean Gaussian with variance $1/N$, where N is the size of the system. (*c*) For an asymmetric SK model with $N = 100$ spins, we plot the probability distribution (color) and fluxes between states (arrows) for simulated time-series at temperatures $T = 0.1$ (left), $T = 1$ (middle), and $T = 10$ (right). In order to visualize the dynamics, the time series are projected onto the first two principal components of the combined data across all three temperatures. The scale is indicated in flux-per-time-step, and the disks represent two-standard-deviation confidence intervals estimated using trajectory bootstrapping (see Methods).

ing simulated dynamics that obey detailed balance and converge to equilibrium (*476*). However, if we relax this constraint to allow for asymmetric interactions, then the system diverges from equilibrium, displaying closed loops of flux between configurations at small scales (Fig. 9.2a). But can these fine-scale violations of detailed balance combine to generate macroscopic non-equilibrium dynamics? To answer this question, we study a system of $N = 100$ spins (matching the 100 parcels in our neuroimaging data), with the interaction between each directed pair of spins drawn independently from a zero-mean Gaussian (Fig. 9.2b). This model is the asymmetric generalization of the Sherrington-Kirkpatrick (SK) model of a spin glass (*607*). After simulating the system at three different temperatures, we perform the same procedure that we applied to the neuroimaging data (Fig. 9.1), projecting the time-series onto the first two principal

components of the combined data and calculating net fluxes in this low-dimensional state space. At high temperature, stochastic fluctuations dominate the system, and we only observe weak fluxes between states (Fig. 9.2c, right). However, as the temperature decreases, the interactions between spins overcome the stochastic fluctuations, giving rise to clear loops of flux (Fig. 9.2c, middle and left). These loops of flux demonstrate that large-scale non-equilibrium dynamics can emerge from fine-scale asymmetries in the interactions between elements. Moreover, by tuning the strength of interactions, a single system can transition from near equilibrium to far from equilibrium, suggesting that the brain may operate at different "effective" temperatures when performing distinct cognitive functions (Fig. 9.1d,e).

9.4 QUANTIFYING ENTROPY PRODUCTION IN COMPLEX SYSTEMS

While fluxes in state space reveal non-equilibrium dynamics, quantifying this non-equilibrium behavior requires measuring the "distance" of a system from equilibrium. One such measure is the entropy production, a central concept in non-equilibrium statistical mechanics (*598*), which quantifies the amount of entropy that a system generates. Consider a system with joint transition probabilities $P_{ij} = \text{Prob}[x_{t-1} = i, x_t = j]$, where x_t is the state of the system at time t. If the dynamics are Markovian (as, for instance, is true for the Ising system), then the entropy production is given by (*562*)

$$S = \sum_{ij} P_{ij} \log \frac{P_{ij}}{P_{ji}}, \tag{9.1}$$

where the sum runs over all states i and j. If the system obeys detailed balance (that is, if $P_{ij} = P_{ji}$ for all pairs of states i and j), then the entropy production vanishes. Conversely, any violation of detailed balance leads to an increase in entropy production, thereby reflecting the non-equilibrium nature of the system.

Calculating the entropy production requires estimating the transition probabilities P_{ij}. However, for complex systems the number of states grows exponentially with the size of the system, making a direct estimate of the entropy production infeasible. To overcome this hurdle, we employ a hierarchical clustering algorithm that groups similar states in our observed data into a single cluster, yielding a reduced number of coarse-grained states (Fig. 9.3a; see Methods). Estimating the entropy production this way yields two desirable properties: First, because a system's entropy production can only decrease with coarse-graining (*208*), in order to establish that a system is fundamentally out of equilibrium, one must simply demonstrate that the coarse-grained entropy production is significantly greater than zero. Second, by defining the clusters hierarchically (*388*), we prove that the estimated entropy production becomes more accurate (ignoring finite data effects) as the number of clusters increases (Fig. 9.9). Indeed, across all temperatures in the Ising system, the estimated entropy production increases with the number of clusters k, thereby providing an improving lower bound on the true entropy production (Fig. 9.3b). Moreover, as the temperature decreases the entropy production

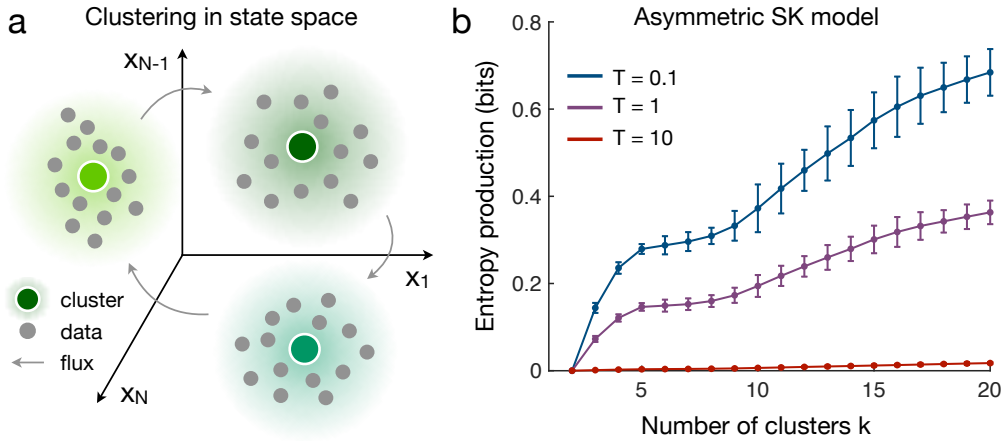


Figure 9.3: Estimating entropy production using hierarchical clustering. (*a*) Schematic of clustering procedure, where axes represent the activities of individual components (e.g., brain regions in the neuroimaging data or spins in the Ising model), points reflect individual states observed in the time-series, shaded regions define clusters (or coarse-grained states), and arrows illustrate possible fluxes between clusters. (*b*) Entropy production in the asymmetric SK model as a function of the number of clusters k for the same time-series studied in Fig. 9.2c, with error bars reflecting two standard deviations estimated using trajectory bootstrapping (see Methods).

grows dramatically, revealing the stark difference in the non-equilibrium nature of the system at high versus low temperature.

9.5 ENTROPY PRODUCTION IN THE BRAIN

We are now prepared to investigate whether the brain operates at different distances from equilibrium when performing distinct functions. We study seven tasks, each of which engages a specific cognitive system: emotional processing, working memory, social inference, language processing, relational matching, gambling, and motor execution (52). To estimate the entropy production, we cluster the neuroimaging data (combined across all subjects and task settings, including rest) into $k = 8$ coarse-grained states, the largest number for which all transitions were observed at least once in each task (Fig. 9.10). Across all tasks and rest, the brain produces a significant amount of entropy (specifically, the entropy production is significantly greater than the noise floor that arises due to finite data; one-sided t-test with $p < 0.001$), confirming that the brain operates out of equilibrium (Fig. 9.4a). Furthermore, the brain produces more entropy during all of the cognitive tasks than at rest, with each task inducing a distinct pattern of fluxes between states (Fig. 9.11). In fact, the motor task (wherein subjects are prompted to perform specific physical movements) induces a 20-fold increase in entropy production over resting-state dynamics, thereby demonstrating that the brain is capable of operating at a wide range of distances from equilibrium.

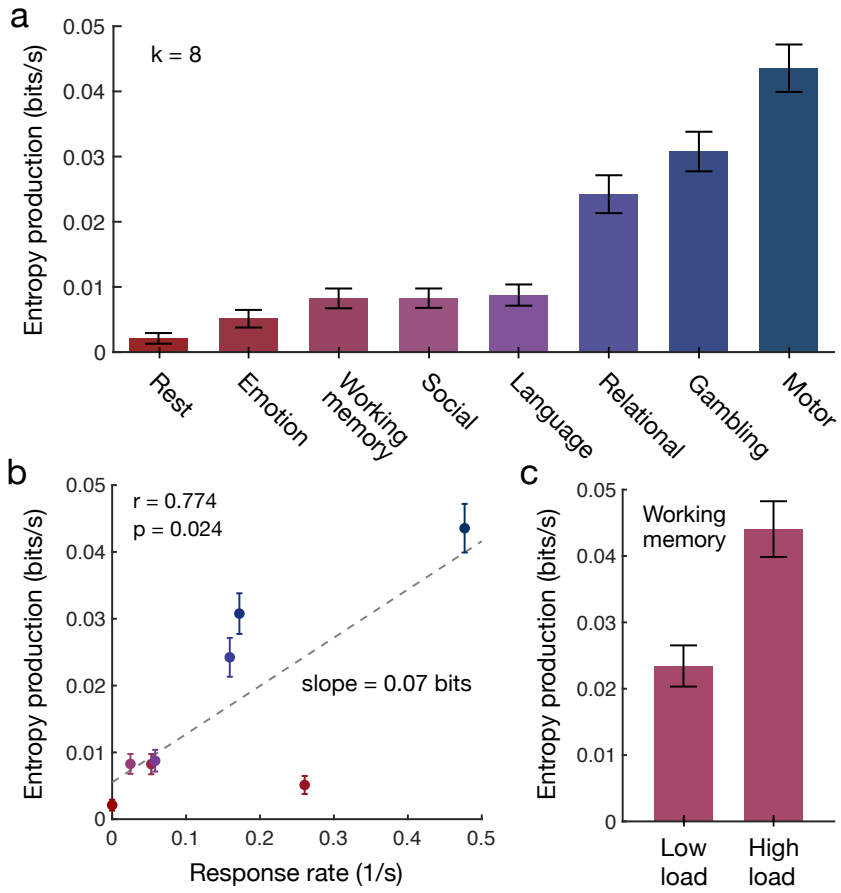


Figure 9.4: Entropy production in the brain varies physical and cognitive demands. (*a*) Entropy production at rest and during seven cognitive tasks, estimated using hierarchical clustering with $k = 8$ clusters. (*b*) Entropy production as a function of response rate (i.e., the frequency with which subjects are asked to physically respond) for the tasks listed in panel (*a*). Each response induces an average 0.07 ± 0.03 bits of produced entropy (Pearson correlation $r = 0.774$, $p = 0.024$). (*c*) Entropy production for low cognitive load and high cognitive load conditions in the working memory task, where low and high loads represent 0-back and 2-back conditions, respectively, in an n-back task. The brain produces significantly more entropy during high-load than low-load conditions (one-sided t-test, $p < 0.001$, $t > 10$, $df = 198$). Across all panels, raw entropy productions (Eq. (9.1)) are divided by the fMRI repetition time $\Delta t = 0.72$ s to compute an entropy production rate, and error bars reflect two standard deviations estimated using trajectory bootstrapping (see Methods).

To carry out the physical and cognitive functions required for each task – from focusing attention to performing computations and responding to stimuli – the brain consumes large amounts of energy (*295*). In living systems generally, such energy consumption is often critical for supporting non-equilibrium dynamics (*255*). Therefore, it is natural to wonder whether increases in physical and cognitive demands drive the brain away from equilibrium. Indeed, across tasks, entropy production increases with the frequency of physical responses (Fig. 9.4b), with each response producing

an additional 0.07 ± 0.03 bits of entropy. Furthermore, within the working memory task (which controls for the frequency of physical responses), the brain produces more entropy during conditions that require greater cognitive effort (Fig. 9.4c). We verify that these findings do not depend on the Markov assumption in Eq. (9.1) (Fig. 9.12), are robust to reasonable variation in the number of clusters k (Fig. 9.13), and cannot be explained by head motion in the scanner (a common confound in fMRI studies (*231*)) nor variance in the neural time-series (Fig. 9.14). Together, these results suggest that physical and cognitive demands, which are supported by the consumption of energy, lead the brain to produce more entropy, thereby driving neural dynamics away from equilibrium.

9.6 CONCLUSIONS

In this study, we describe a method for investing non-equilibrium dynamics by quantifying the amount of entropy that a system produces in its environment. While microscopic non-equilibrium processes are known to be vital for molecular and cellular operations (*103*, *197*, *328*, *389*, *441*, *641*, *718*), here we show that non-equilibrium dynamics also arise at large scales in complex living systems. Analyzing whole-brain imaging data, we find not only that the human brain functions out of equilibrium, but that the brain's entropy production (that is, its distance from equilibrium) increases with physical and cognitive exertion. Notably, the tools presented are non-invasive, applying to any system with time-series data, and can be used to study stochastic steady-state dynamics, rather than deterministic dynamics that trivially break detailed balance. Furthermore, the framework is not limited to the brain, but instead can be applied broadly to probe for non-equilibrium dynamics in complex systems, including collective behavior in human and animal populations (*126*), correlated patterns of neuronal firing (*501*), and aggregated activity in molecular and cellular networks (*380*, *676*).

9.7 METHODS

9.7.1 *Calculating fluxes*

Consider time-series data gathered in a time window t_{tot}, and let n_{ij} denote the number of observed transitions from state i to state j. The flux rate from state i to state j is given by $\omega_{ij} = (n_{ij} - n_{ji})/t_{tot}$. For the flux currents in Figs. 9.1d,e and 9.2c, the states of the system are points (x, y) in two-dimensional space, and the state probabilities are estimated by $p(x, y) = t_{(x,y)}/t_{tot}$, where $t_{(x,y)}$ is the time spent in state (x, y). The magnitude and direction of the flux through a given state (x, y) is defined by the flux vector (*67*)

$$\mathbf{u}(x, y) = \frac{1}{2}\begin{pmatrix} \omega_{(x-1,y),(x,y)} + \omega_{(x,y),(x+1,y)} \\ \omega_{(x,y-1),(x,y)} + \omega_{(x,y),(x,y+1)} \end{pmatrix}. \tag{9.2}$$

In a small number of cases, two consecutive states in the observed time-series $\mathbf{x}(t) = (x(t), y(t))$ and $\mathbf{x}(t+1) = (x(t+1), y(t+1))$ are not adjacent in state space. In these cases, we perform a linear interpolation between $\mathbf{x}(t)$ and $\mathbf{x}(t+1)$ in order to calculate the fluxes between adjacent states.

9.7.2 *Estimating errors using trajectory bootstrapping*

The finite length of time-series data limits the accuracy with which quantities can be estimated. In order to calculate error bars on all estimated quantities, we apply trajectory bootstrapping (*67*, *604*). We first record the list of transitions

$$I = \begin{pmatrix} i_1 & i_2 \\ i_2 & i_3 \\ \vdots & \vdots \\ i_{L-1} & i_L \end{pmatrix}, \tag{9.3}$$

where i_ℓ is the ℓ^{th} state in the time-series, and L is the length of the time-series. From the transition list I, one can calculate all of the desired quantities; for instance, the fluxes are estimated by

$$\omega_{ij} = \frac{1}{t_{\text{tot}}} \sum_\ell \delta_{i,I_{\ell,1}} \delta_{j,I_{\ell,2}} - \delta_{j,I_{\ell,1}} \delta_{i,I_{\ell,2}}. \tag{9.4}$$

We remark that when analyzing the neural data, although we concatenate the time-series across subjects, we only include transitions in I that occur within the same subject. That is, we do not include the transitions between adjacent subjects in the concatenated time-series.

To calculate errors, we construct bootstrap trajectories (of the same length L as the original time-series) by sampling the rows in I with replacement. For example, to compute errors for the flux vectors $\mathbf{u}(\mathbf{x})$ in Figs. 9.1d,e and 9.2c, we first estimate the covariance matrix $\text{Cov}(u_1(\mathbf{x}), u_2(\mathbf{x}))$ by averaging over bootstrapped trajectories. Then, for each flux vector, we visualize its error by plotting an ellipse with axes aligned with the eigenvectors of the covariance matrix and radii equal to twice the square root of the corresponding eigenvalues (Fig. 9.5). All errors throughout the manuscript are calculated using 100 bootstrap trajectories.

The finite data length also induces a noise floor for each quantity, which is present even if the temporal order of the time-series is destroyed. To estimate the noise floor, we construct bootstrap trajectories by sampling individual data points from the time-series. We contrast these bootstrap trajectories with those used to estimate errors above, which preserve transitions by sampling the rows in I. The noise floor, which is calculated for each quantity by averaging over the bootstrap trajectories, is then compared with the estimated quantities. For example, rather than demonstrating that the average entropy

productions in Fig. 9.4a are greater than zero, we establish that the distribution over entropy productions is significantly greater than the noise floor using a one-sided *t*-test with $p < 0.001$.

9.7.3 *Simulating the asymmetric Ising model*

The asymmetric Ising model is defined by a (possibly asymmetric) interaction matrix J, where $J_{\alpha\beta}$ represents the influence of spin β on spin α (Fig. 9.2a), and a temperature $T \geqslant 0$ that tunes the strength of stochastic fluctuations. Here, we study a system with $N = 100$ spins, where each directed interaction $J_{\alpha\beta}$ is drawn independently from a zero-mean Gaussian with variance $1/N = 0.01$ (Fig. 9.2b). One can additionally include external fields h_α, but for simplicity here we set them to zero. The state of the system is defined by a vector $\mathbf{x} = (x_1, \ldots, x_N)$, where $x_\alpha = \pm 1$ is the state of spin α. To generate time series, we employ Glauber dynamics with synchronous updates, a common Monte Carlo method for simulating Ising systems (*476*). Specifically, given the state of the system $\mathbf{x}(t)$ at time t, the probability of spin α being "up" at time $t+1$ (that is, the probability that $x_\alpha(t+1) = 1$) is given by

$$\text{Prob}[x_\alpha(t+1) = 1] = \frac{\exp\left(\frac{1}{T}\sum_\beta J_{\alpha\beta} x_\beta(t)\right)}{\exp\left(\frac{1}{T}\sum_\beta J_{\alpha\beta} x_\beta(t)\right) + \exp\left(-\frac{1}{T}\sum_\beta J_{\alpha\beta} x_\beta(t)\right)}. \tag{9.5}$$

Stochastically updating each spin α according to Eq. (9.5), one arrives at the new state $\mathbf{x}(t+1)$. For each temperature in the Ising calculations in Figs. 9.2c and 9.3b, we generate a different time-series of length $L = 100,000$ with $10,000$ trials of burn-in.

9.7.4 *Hierarchical clustering*

To estimate the entropy production of a system, one must first calculate the transition probabilities $P_{ij} = n_{ij}/(L-1)$. For complex systems, the number of states i (and therefore the number of transitions $i \to j$) grows exponentially with the size of the system N. For example, in the Ising model each spin α can take one of two values ($x_\alpha = \pm 1$), leading to 2^N possible states and 2^{2N} possible transitions. In order to estimate the transition probabilities P_{ij}, one must observe each transition $i \to j$ at least once, which requires significantly reducing the number of states in the system. Rather than defining coarse-grained states *a priori*, complex systems (and the brain in particular) often admit natural coarse-grained descriptions that are uncovered through dimensionality-reduction techniques (*156*, *165*, *408*).

Although one can use any coarse-graining technique to implement our framework and estimate entropy production, here we employ hierarchical k-means clustering for two reasons: (i) Generally, k-means is perhaps the most common and simplest clustering algorithm, with demonstrated effectiveness fitting neural dynamics (*156*, *408*); and (ii) specifically, by defining the clusters hierarchically we prove that the estimated entropy

production becomes more accurate as the number of clusters increases (ignoring finite data effects; Fig. 9.9).

In k-means clustering, one begins with a set of states (for example, those observed in our time-series) and a number of clusters k. Each observed state x is randomly assigned to a cluster i, and one computes the centroid of each cluster. On the following iteration, each state is re-assigned to the cluster with the closest centroid (here we use cosine similarity to determine distance). This process is repeated until the cluster assignments no longer change. In a hierarchical implementation, one begins with two clusters; then one cluster is selected (typically the one with the largest spread in its constituent states) to be split into two new clusters, thereby defining a total of three clusters. This iterative splitting is continued until one reaches the desired number of clusters k. In Sec. 9.8.5, we show that hierarchical clustering provides an increasing lower-bound on the entropy production; and in Sec. 9.8.6, we demonstrate how to choose the number of clusters k.

9.7.5 *Neural data*

The whole-brain dynamics used in this study are measured and recorded using blood-oxygen-level-dependent (BOLD) functional magnetic resonance imagining (fMRI) collected from 590 healthy adults as part of the Human Connectome Project (*52*, *678*). BOLD fMRI estimates neural activity by calculating contrasts in blood oxygen levels, without relying on invasive injections and radiation (*543*). Specifically, blood oxygen levels (reflecting neural activity) are measured within three-dimensional non-overlapping voxels, spatially contiguous collections of which each represent a distinct brain region (or parcel). Here, we consider a parcellation that divides the cortex into 100 brain regions that are chosen to optimally capture the functional organization of the brain (*662*). After processing the signal to correct for sources of systematic noise such as head motion (see Sec. 9.8.11), the activity of each brain region is discretized in time, yielding a time-series of neural activity. For each subject, the shortest scan (corresponding to the emotional processing task) consists of 176 discrete measurements. In order to control for variability in data size across tasks, for each subject we only study the first 176 measurements in each task.

9.8 SUPPLEMENTARY MATERIAL

In this Supplementary material, we provide extended analysis and discussion to support the results presented in the main text. In Sec. 9.8.1, we describe how the flux vectors (in Figs. 1d,e and 2c of the main text) are calculated and illustrated. In Sec. 9.8.2, we show that principal component analysis (PCA) provides a natural low-dimensional embedding of neural dynamics that we can use to visualize fluxes between brain states. In Sec. 9.8.3, we show that, although the brain functions out of equilibrium, it does operate at a steady state. Demonstrating that the brain operates at a non-equilibrium steady-state opens the door for future investigations using tools and intuitions that have recently been generalized from traditional statistical mechanics (*187*, *211*, *599*). In Sec. 9.8.4, we show that if one shuffles the order of neural time-series data (thereby destroying the arrow of time), then the fluxes between states vanish and the system returns to equilibrium. In Sec. 9.8.5, we establish that estimating entropy production using hierarchical clustering yields two desirable properties: First, because a system's entropy production can only decrease with coarse-graining (*208*), in order to establish that a system is fundamentally out of equilibrium, one must simply demonstrate that the coarse-grained entropy production is significantly greater than zero. Second, by defining the clusters hierarchically (*338*), we prove that the estimated entropy production becomes more accurate as the number of clusters increases. In Sec. 9.8.6, we demonstrate how to choose the number of clusters (or coarse-grained states) k when estimating the entropy production. In Sec. 9.8.7, we present the flux between coarse-grained states in the neural dynamics as directed networks, which we refer to as flux networks. We demonstrate that these flux networks vary in structure across different cognitive tasks. In Secs. 9.8.8-9.8.10, we show that the entropy production results in Fig. 9.4 do not depend on the assumption that the neural dynamics are Markovian (Sec. 9.8.8), are robust to reasonable variation in the number of coarse-grained states k (Sec. 9.8.9), and cannot be explained by head movement within the scanner nor variance in the neural time-series (Sec. 9.8.10). Finally, in Sec. 9.8.11, we detail how the neural data was processed.

9.8.1 *Visualizing flux currents*

In order to visualize net fluxes in neural dynamics, we project the dynamics onto the first two principal components and employ a technique known as probability flux analysis (*67*). The net flux of transitions from a given state (x, y) to its neighboring states can be visualized using the flux vector

$$\mathbf{u}(x, y) = \frac{1}{2}\begin{pmatrix} \omega_{(x-1,y),(x,y)} + \omega_{(x,y),(x+1,y)} \\ \omega_{(x,y-1),(x,y)} + \omega_{(x,y),(x,y+1)} \end{pmatrix}. \tag{9.6}$$

To compute the errors for a given flux vector $\mathbf{u}(\mathbf{x})$, we calculate the covariance matrix $\mathrm{Cov}(u_1(\mathbf{x}), u_2(\mathbf{x}))$ by averaging over 100 bootstrapped trajectories. Then, we illustrate

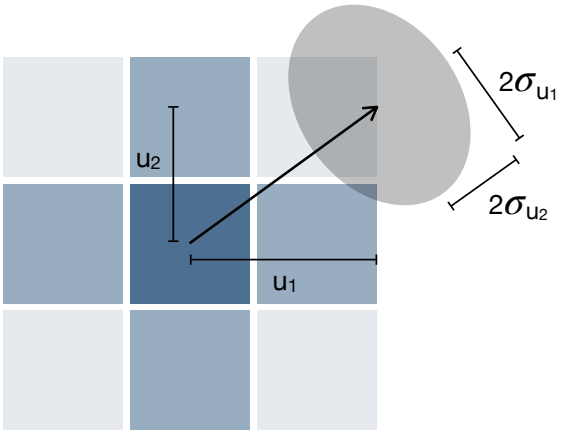


Figure 9.5: Visualizing flux vectors. Schematic demonstrating how we illustrate the flux of transitions through a state (vector) and the errors in estimating the flux (ellipse).

the errors by plotting an ellipse whose axes are aligned with the eigenvectors of the covariance matrix and whose radii are equal to twice the square root of the corresponding eigenvalues (Fig. 9.5).

9.8.2 *Low-dimensional embedding using PCA*

In order to visualize net fluxes between states in a complex system, we must project the dynamics onto two dimensions. While any pair of dimensions can be used to probe for broken detailed balance, a natural choice is the first two principal components of the time-series data. Indeed, principal component analysis has been widely used to uncover low-dimensional embeddings of large-scale neural dynamics (*165*, *230*, *610*). Combining the time-series data from the rest and gambling task scans (that is, the data studied in Fig. 9.1), we find that the first two principal components capture over 30% of the total variance in the observed recordings (Fig. 9.6a), thereby comprising a natural choice for two-dimensional projections. Moreover, we confirm that the projected dynamics capture approximately the same amount of variance in both the rest and gambling tasks, confirming that PCA is not overfitting the neural dynamics in one task or another (Fig. 9.6b).

9.8.3 *The brain operates at a stochastic steady state*

Some of the tools and intuitions developed in traditional statistical mechanics to study equilibrium systems have recently been generalized to systems that exist at non-equilibrium steady states (*599*). For example, Evans *et al.* generalized the second law of thermodynamics to non-equilibrium steady-state systems by discovering the (steady state) fluctuation theorem (*211*). More recently, Dieterich *et al.* showed that, by mapping their dynamics to an equilibrium system at an effective temperature,

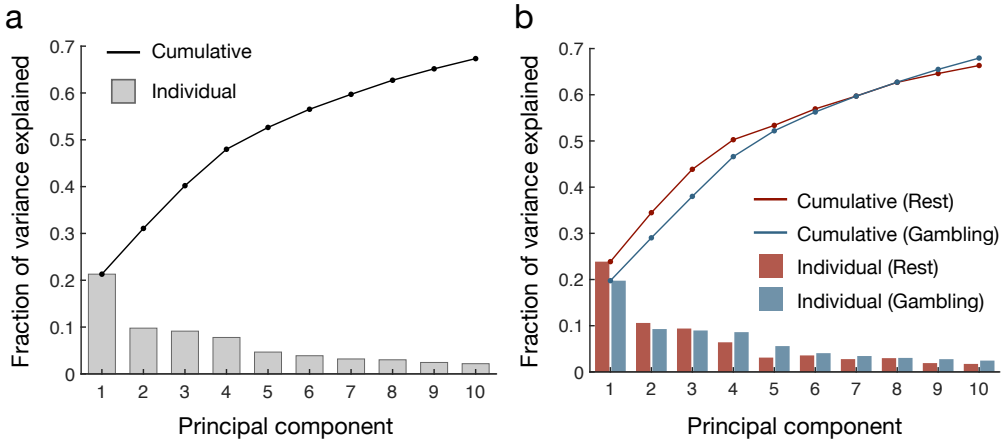


Figure 9.6: PCA reveals low-dimensional embedding of neural dynamics. (*a*) Fraction of variance explained by first ten principal components (line) and increase in explained variance for each principal component (bars) in the combined rest and gambling data. (*b*) For the same principal components (calculated for the combined rest and gambling data), we plot the fraction of variance explained (lines) and individual increases in explained variance (bars) for the rest (red) and gambling (blue) data.

some non-equilibrium steady-state systems are governed by a generalization of the fluctuation-dissipation theorem (*187*). Thus, it is both interesting and practical to investigate whether the brain operates at a non-equilibrium steady state.

We establish in the main text that the brain operates out of equilibrium. To determine if the brain functions at a steady state, we must examine whether its state probabilities are stationary in time; that is, letting p_i denote the probability of state i, we must determine whether $\dot{p}_i = dp_i/dt = 0$ for all states i. The change in the probability of a state is equal to the net rate at which transitions flow into versus out of a state. For the two-dimensional dynamics studied in Fig. 1 in the main text, this relation takes the form

$$\frac{dp_{(x,y)}}{dt} = \omega_{(x-1,y),(x,y)} - \omega_{(x,y),(x+1,y)} + \omega_{(x,y-1),(x,y)} - \omega_{(x,y),(x,y+1)}, \tag{9.7}$$

where $\omega_{ij} = (n_{ij} - n_{ji})/t_{tot}$ is the flux rate from state i to state j, n_{ij} is the number of observed transitions $i \to j$, and t_{tot} is the temporal duration of the time-series (*67*).

Here, we calculate the changes in state probabilities for both the rest and gambling scans. Across all states in both task settings, we find that these changes are indistinguishable from zero when compared to statistical noise (Fig. 9.7). Specifically, the the changes in state probabilities are much less than twice their standard deviations, indicating that they cannot be significantly distinguished from zero with a p-value less than 0.05. Combined with the results from the main text, the stationarity of the neural state probabilities demonstrates that the brain operates at a non-equilibrium steady-state.

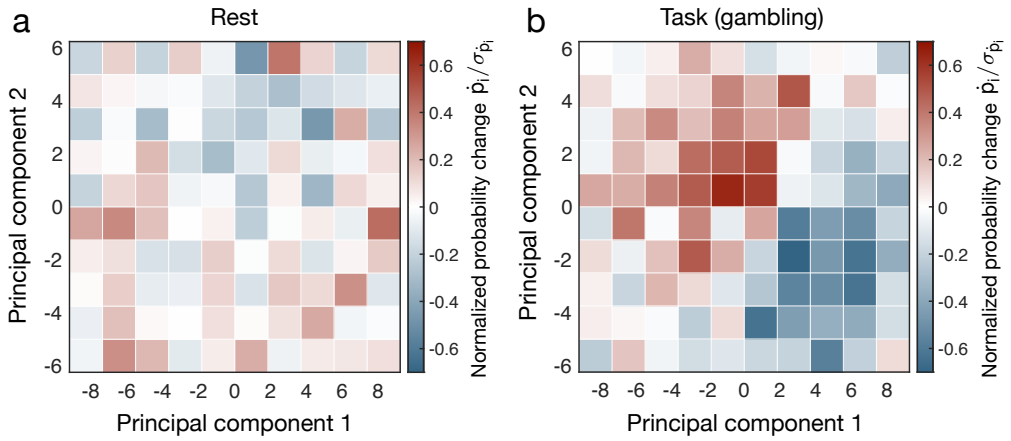


Figure 9.7: Small changes in state probabilities imply steady-state dynamics. Change in state probabilities $\dot{p}_i$, normalized by the standard deviation $\sigma_{\dot{p}_i}$, plotted as a function of the first two principal components at rest (*a*) and during the gambling task (*b*).

9.8.4 *Shuffling time-series restores equilibrium*

In the main text, we demonstrate that the brain operates out of equilibrium by exhibiting net fluxes between states (Fig. 9.1d,e in the main text). These fluxes break detailed balance and establish an arrow of time. Here we demonstrate that if the arrow of time is destroyed by shuffling the order of the neural time-series, then the fluxes vanish and equilibrium is restored. Specifically, for both the rest and gambling task scans, we generate 100 surrogate time-series with the order of the data randomly shuffled. Averaging across these shuffled time-series, we find that the fluxes between states are vanishingly small compared to statistical noise (Fig. 9.8), thus illustrating that the system has returned to equilibrium.

9.8.5 *Bounding entropy production using hierarchical clustering*

Complex systems are often high-dimensional, with the number of possible states or configurations growing exponentially with the size of the system. In order to estimate the entropy production of a complex system, we must reduce the number of states through the use of coarse-graining, or dimensionality reduction, techniques. Interestingly, the entropy production admits a number of strong properties under coarse-graining (*208*, *264*, *532*, *611*). Of particular interest is the fact that the entropy production can only decrease under coarse-graining (*208*). Specifically, given two descriptions of a system, a "microscopic" description with states $\{i\}$ and a "macroscopic" description with states $\{i'\}$, we say that the second description is a coarse-graining of the first if there exists a surjective map from the microstates $\{i\}$ to the macrosctates $\{i'\}$ (that is, if each microstate i gets mapped to a unique macrostate i'; Fig. 9.9a). Given such a coarse-graining, Esposito showed (*208*) that the entropy production of the macroscopic description S' can be no larger than that of the microscopic description S;

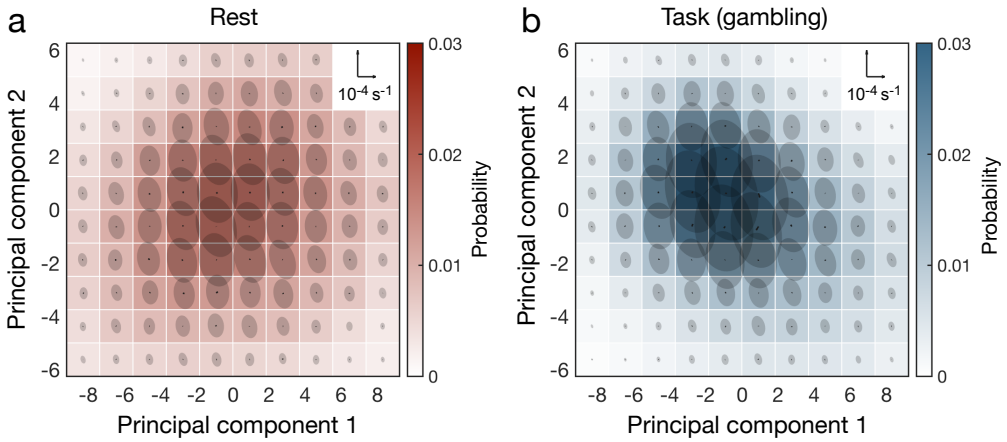


Figure 9.8: Shuffled data do not exhibit net fluxes between brain states. Probability distribution (color) and nearly imperceivable fluxes between states (arrows) for neural dynamics, which are shuffled and projected onto the first two principal components, both at rest (*a*) and during a gambling task (*b*). The flux scale is indicated in the upper right, and the disks represent two-standard-deviation confidence intervals for fluxes estimated using trajectory bootstrapping (see Methods).

in other words, the coarse-grained entropy production provides a lower bound for the original value, such that $S' \leqslant S$.

The monotonic decrease of the entropy production under coarse-graining implies two desirable mathematical results. First, if one finds that any coarse-grained description of a system is out of equilibrium (that is, if the coarse-grained entropy production is significantly greater than zero), then one has immediately established that the full microscopic system is out of equilibrium (since the "true" microscopic entropy production is at least as large as the coarse-grained value). We use this fact in the main text to show – only by studying coarse-grained dynamics – that the brain fundamentally operates far from equilibrium.

Second, here we show that hierarchical clustering provides a hierarchy of lower bounds on the true entropy production. In hierarchical clustering, each cluster (or coarse-grained state) at one level of description (with k clusters) maps to a unique cluster at the level below (with $k-1$ clusters; Fig. 9.9b). This process can either be carried out by starting with a large number of clusters and then iteratively picking pairs of clusters to combine (known as agglomerative clustering), or by starting with a small number of clusters and then iteratively picking one cluster to split into two (known as divisive clustering, which we employ in our analysis) (*338*). In both cases, the mapping from k clusters to $k-1$ clusters is surjective, thereby defining a coarse-graining of the system. Thus, letting $S^{(k)}$ denote the entropy production estimated with k clusters, hierarchical clustering defines a hierarchy of lower bounds on the true entropy production S:

$$0 = S^{(1)} \leqslant S^{(2)} \leqslant S^{(3)} \leqslant \ldots \leqslant S. \tag{9.8}$$

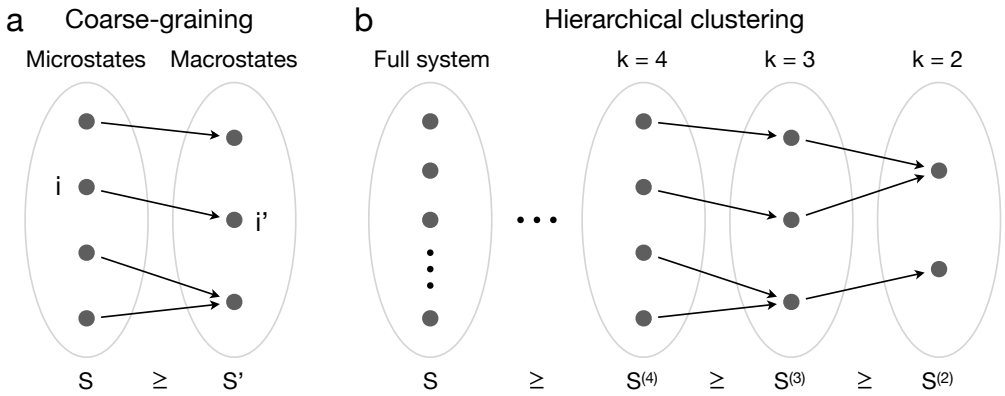


Figure 9.9: Hierarchy of lower bounds on the entropy production. (*a*) Coarse-graining is defined by a surjective map from a set of microstates $\{i\}$ to a set of macrostates $\{i'\}$. Under coarse-graining the entropy production can only decrease or remain the same. (*b*) In hierarchical clustering, states are iteratively combined to form new coarse-grained states (or clusters). Each iteration defines a coarse-graining from k states to $k-1$ states, thereby forming a hierarchy of lower bounds on the entropy production.

This hierarchy, in turn, demonstrates that the estimated entropy production $S^{(k)}$ becomes more accurate with increasing k.

We remark that the discussion above neglects finite data effects. We recall that estimating the entropy production requires first estimating the transition probabilities P_{ij} from state i to state j. This means that for k clusters, one must estimate k^2 different probabilities. Thus, while increasing k improves the accuracy of the estimated entropy production in theory, in practice increasing k eventually leads to sampling issues that decrease the accuracy of the estimate. Given these competing influences, when analyzing real data the goal should be to choose k such that it is as large as possible while still providing accurate estimates of the transition probabilities. We discuss how to choose k in a reasonable manner in the following section.

9.8.6 *Choosing the number of coarse-grained states*

As discussed above, when calculating the entropy production, we wish to choose a number of coarse-grained states k that is as large as possible while still arriving at an accurate estimate of the transition probabilities. One simple condition for estimating each transition probability P_{ij} is that we observe the transition $i \to j$ at least once in the time-series. For all of the different tasks, Fig. 9.10a shows the fraction of the k^2 state transitions that are left unobserved after coarse-graining with k clusters. We find that $k = 8$ is the largest number of clusters for which the fraction of unobserved transitions equals zero (within statistical errors) for all tasks; that is, the largest number of clusters for which all state transitions across all tasks were observed at least once. This is the primary reason why we used $k = 8$ coarse-grained states to analyze the brain's entropy production in the main text (Fig. 9.4).

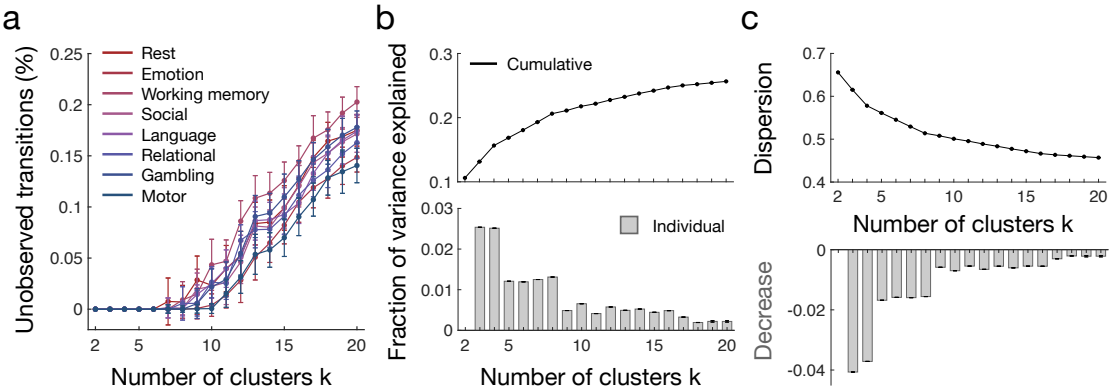


Figure 9.10: Choosing the number of coarse-grained states k. (*a*) Fraction of the k^2 state transitions that remain unobserved after hierarchical clustering with k clusters for the different tasks. Error bars represent two standard deviations over 100 bootstrap trajectories for each task. (*b*) Percent variance explained (top) and the increase in explained variance from $k-1$ to k clusters (bottom) as functions of k. (*c*) Dispersion, or the average distance between data points within a cluster (top), and the decrease in dispersion from $k-1$ to k clusters (bottom) as functions of k.

Interestingly, we find that $k = 8$ coarse-grained states is a good choice for two additional reasons. The first comes from studying the amount of variance explained by k clusters (Fig. 9.10b). We find that the increase in explained variance from $k-1$ to k clusters is roughly constant for $k = 3$ and 4, then $k = 5$ to 8, and then $k = 9$ to 16. This pattern means that $k = 4$, 8, and 16 are natural choices for the number of coarse-grained states, since any further increase (say from $k = 8$ to 9) will yield a smaller improvement in explained variance. Similarly, the second reason for choosing $k = 8$ comes from studying the average distance between states within a cluster, which is known as the dispersion (Fig. 9.10c). Intuitively, a coarse-grained description with low dispersion provides a good fit of the observed data. Similar to the explained variance, we find that the decrease in dispersion from $k-1$ to k clusters is nearly constant for $k = 3$ to 4, then $k = 5$ to 8, and then $k = 9$ to 16, once again suggesting that $k = 4$, 8, and 16 are natural choices for the number of clusters. Together, these results demonstrate that the coarse-grained description with $k = 8$ states provides a good fit to the neural time-series data while still allowing for an accurate estimate of the entropy production in each task.

9.8.7 *Flux networks: Visualizing flux between coarse-grained states*

In Fig. 9.4, we demonstrated that the brain has the capacity to operate at a wide range of distances from equilibrium. We did so by estimating the amount of entropy the brain produces during different cognitive tasks. In addition to investigating the entropy production, one can also examine the specific neural processes underlying the brain's non-equilibrium behavior, which are encoded in the fluxes between coarse-grained states.

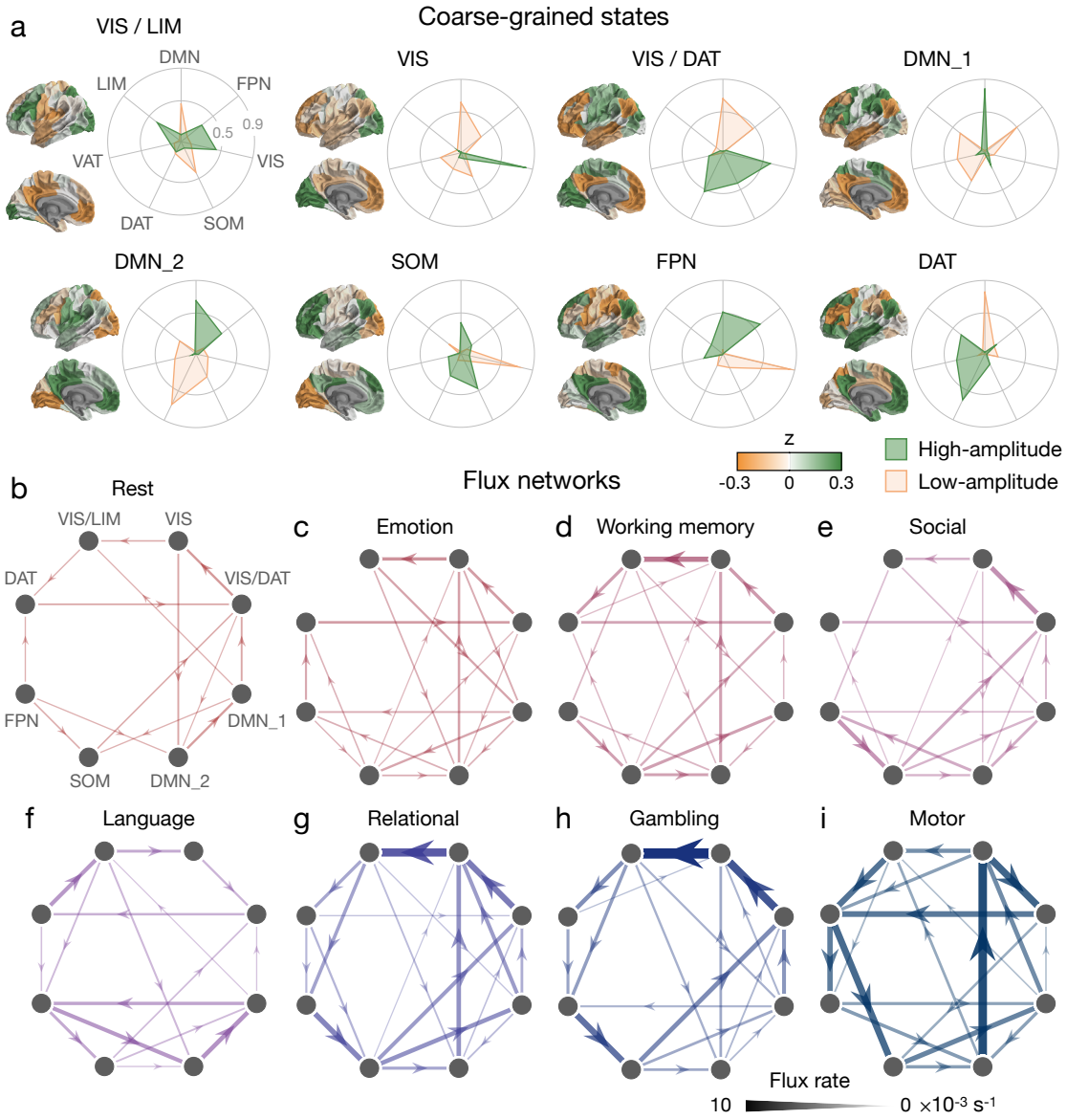


Figure 9.11: Flux networks reveal non-equilibrium dynamics unique to each cognitive task. (*a*) Coarse-grained brain states calculated using hierarchical clustering (k = 8), with surface plots indicating the z-scored activation of different brain regions. For each state, we calculate the cosine similarity between its high-amplitude (green) and low-amplitude (orange) components and seven pre-defined neural systems (*662*): default mode (DMN), frontoparietal (FPN), visual (VIS), somatomotor (SOM), dorsal attention (DAT), ventral attention (VAT), and limbic (LIM). We label each state based on its largest high-amplitude cosine similarities. (*b-i*) Flux networks illustrating the fluxes between the eight coarse-grained states at rest (*b*) and during seven cognitive tasks: emotional processing (*c*), working memory (*d*), social inference (*e*), language processing (*f*), relational matching (*g*), gambling (*h*), and motor execution (*i*). Edge weights indicate flux rates, and fluxes are only included if they are significant relative to the noise floor induced by the finite data length (one-sided *t*-test, $p < 0.001$).

We find that each of the $k = 8$ states corresponds to high-amplitude activity in one or two cognitive systems (*662*) (Fig. 9.11a). For each task, we can visualize the pattern of fluxes as a network, with nodes representing the coarse-grained states and directed edges reflecting net fluxes between states (Fig. 9.11b-i). These flux networks illustrate, for example, that the brain nearly obeys detailed balance during rest (Fig. 9.11b). Interestingly, in the emotion, working memory, social, relational, and gambling tasks (Fig. 9.11c-e,g,h) – all of which involve visual stimuli – the strongest fluxes connect visual (VIS) states. By contrast, these fluxes are weak in the language task (Fig. 9.11f), which only involves auditory stimuli. Finally, in the motor task, wherein subjects are prompted to make physical movements, the dorsal attention (DAT) state mediates fluxes between disparate parts of the network (Fig. 9.11i), perhaps reflecting the role of the DAT system in directing goal-oriented attention (*226*, *691*). In this way, the brain's non-equilibrium dynamics are not driven by a single underlying mechanism, but rather emerge from a complex pattern of fluxes that changes depending on the task. Examining the structural properties and cognitive neuroscientific interpretations of these flux networks is an important direction for future studies.

9.8.8 *Testing the Markov assumption*

In the main text, we employ a definition of entropy production that relies on the assumption that the time-series is Markovian; that is, that the state x_t of the system at time t depends only on the previous state x_{t-1} at time $t-1$. Specifically, the entropy production of a Markov system is given by

$$S = \sum_{ij} P_{ij} \log \frac{P_{ij}}{P_{ji}}, \tag{9.9}$$

where $P_{ij} = \text{Prob}[x_{t-1} = i, x_t = j]$ is the probability of observing the transition $i \to j$. For real time-series data, however, the dynamics may not be Markovian, and Eq. (9.9) is not exact. In general, the entropy production (per trial) is given by (*562*, *563*)

$$S = \lim_{\ell \to \infty} \frac{1}{\ell} \sum_{i_1,\ldots,i_{\ell+1}} P_{i_1,\ldots,i_{\ell+1}} \log \frac{P_{i_1,\ldots,i_{\ell+1}}}{P_{i_{\ell+1},\ldots,i_1}}, \tag{9.10}$$

where $P_{i_1,\ldots,i_{\ell+1}} = \text{Prob}[x_{t-\ell} = i_1, \ldots, x_t = i_{\ell+1}]$ is the probability of observing the sequence of states $i_1, \ldots, i_{\ell+1}$. If the dynamics are Markovian, for example, then the limit converges for $\ell = 1$ and we recover Eq. (9.9) (*562*). In general, one can approximate Eq. (9.10) by evaluating the function inside the limit for ℓ as large as possible. In order to do so, however, one must estimate $k^{\ell+1}$ different probabilities for a system with k states. Thus, given data limitations, it is often impractical to estimate the entropy production beyond the Markov approximation ($\ell = 1$).

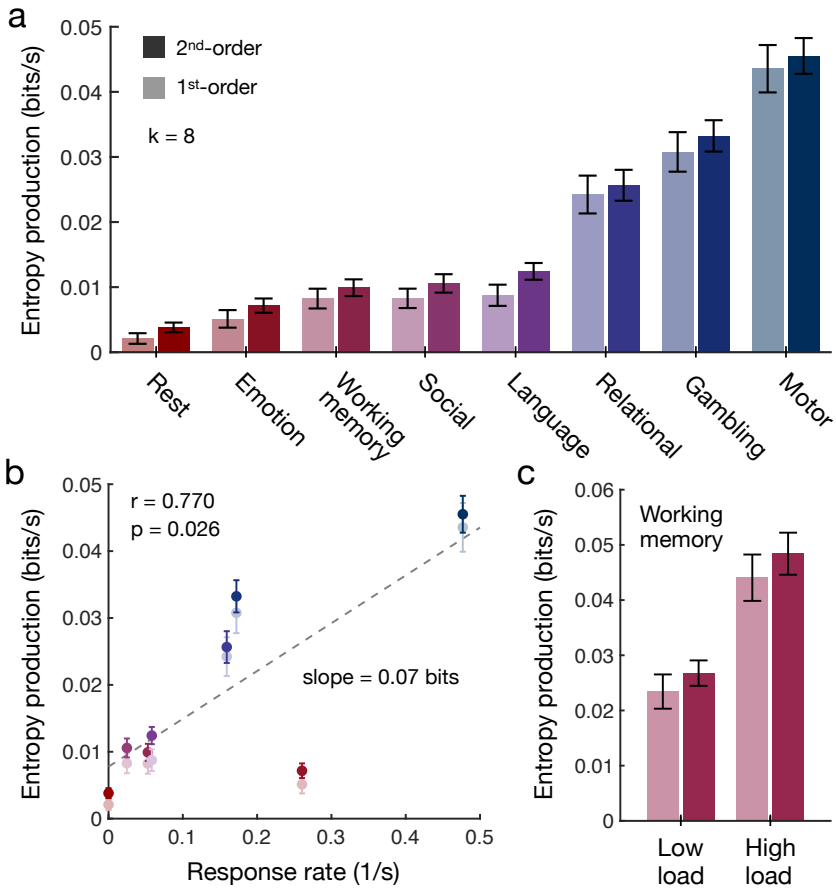


Figure 9.12: Second-order approximation of entropy production in the brain. (*a*) Second-order entropy production at rest and during seven cognitive tasks (dark bars), estimated using hierarchical clustering with $k = 8$ clusters. For comparison, we also include the first-order entropy productions from Fig. 9.4a (light bars). (*b*) Second-order entropy production as a function of response rate for the tasks listed in panel (*a*) (dark points). Each response induces an average 0.07 ± 0.03 bits of produced entropy (Pearson correlation $r = 0.770$, $p = 0.026$). For comparison, we include the first-order entropy productions from Fig. 9.4b (light points). (*c*) We find a significant difference in the second-order entropy production between low cognitive load and high cognitive load conditions in the working memory task (dark bars), where low and high loads represent 0-back and 2-back conditions, respectively (one-sided t-test, $p < 0.001$, $t > 10$, $df = 198$). For comparison, we include the first-order entropy productions from Fig. 9.4c (light bars). Across all panels, second-order entropy productions (calculated using Eq. (9.11)) are divided by the fMRI repetition time $\Delta t = 0.72$ s to compute an entropy production rate, and error bars reflect two standard deviations estimated using trajectory bootstrapping (see Methods).

Here we demonstrate that the main conclusions about entropy production in the brain (summarized in Fig. 9.4) do not depend qualitatively on the Markov approximation in Eq. (9.9). To do so, we consider the second-order approximation

$$S \approx \frac{1}{2} \sum_{i,j,k} P_{ijk} \log \frac{P_{ijk}}{P_{kji}}, \tag{9.11}$$

which incorporates information about sequences of length three. As in the main text, we cluster the neural data using $k = 8$ coarse-grained states. Given that we are now required to estimate $k^3 = 512$ probabilities rather than just $8^2 = 64$, there are inevitably entries in the sum in Eq. (9.11) that are infinite (i.e., those corresponding to reverse-time sequences $k \to j \to i$ that are not observed in the time-series). As is common (*562*, *563*), we simply ignore these terms.

Across the different task settings, we find that the second-order entropy productions are nearly identical to the first-order (Markov) approximations presented in the main text (Fig. 9.12a). Moreover, the second-order entropy production remains significantly correlated with the frequency of physical responses in different tasks, with each response still inducing an additional 0.07 ± 0.03 bits of produced entropy (Fig. 9.12b). Finally, in the working memory task, the second-order entropy production remains larger for high-load conditions than low-load conditions (Fig. 9.12c), suggesting that cognitive demands drive the brain away from equilibrium. Together, these results demonstrate that the brain's entropy production is well-approximated by the Markov formulation used in the main text (Eq. (9.9)).

9.8.9 *Varying the number of coarse-grained states*

In Sec. 9.8.6, we presented methods for choosing the number of coarse-grained states k, concluding that $k = 8$ is an appropriate choice for our neural data. However, it is important to check that the entropy production results (summarized in Fig. 9.4) do not vary significantly with our choice of k. In Fig. 9.13a, we plot the estimated entropy production for each task setting (including rest) as a function of the number coarse-grained states k. We find that the tasks maintain approximately the same ordering across all choices for k considered, with the brain producing the least entropy during rest, the most entropy during the motor task, and the second most entropy during the gambling task. Furthermore, we find that the correlation between entropy production and physical response rate (Fig. 9.4b) remains significant for all $k \leqslant 8$ (that is, for all choices of k for which we observe all transitions at least once in each task; Fig. 9.10a) as well as $k = 9$, 11, 12, 13, and 14 (Fig. 9.13b). We remark that we do not study the case $k = 2$ because the entropy production is zero by definition for two-state systems (Fig. 9.13a). Finally, we confirm that the brain produces significantly more entropy during high-cognitive-load conditions than low-cognitive-load conditions in the working memory task (Fig. 9.4c) for all choices of k considered (Fig. 9.13c). Together, these results demonstrate that the relationships between entropy production and physical and cognitive effort are robust to reasonable variation in the number of coarse-grained states k.

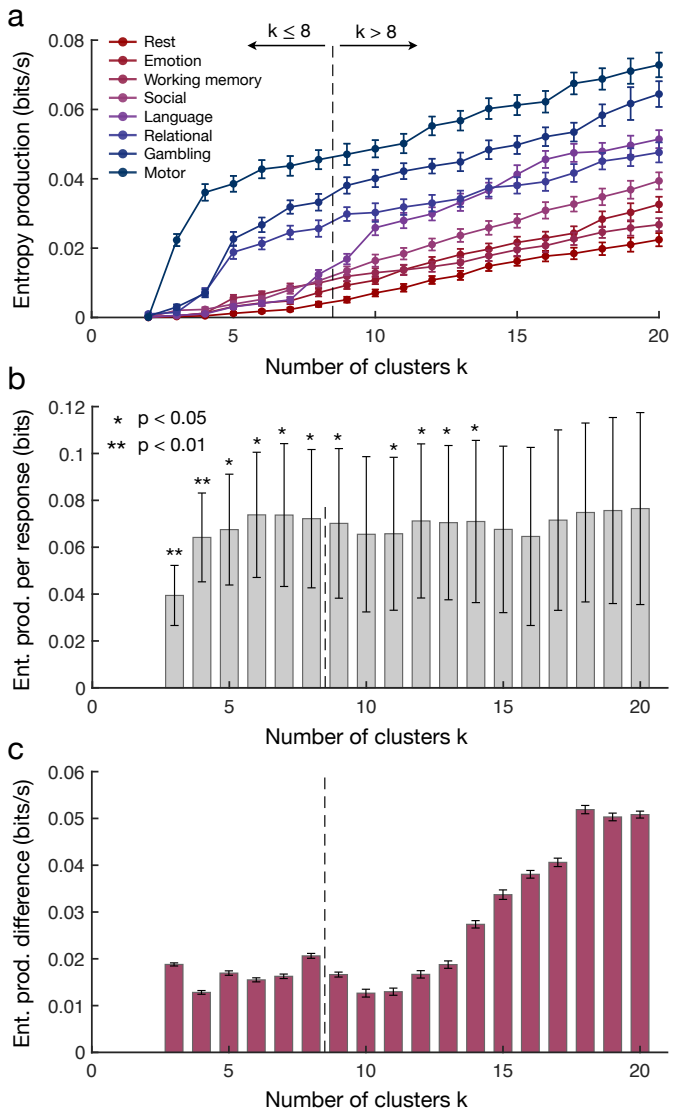


Figure 9.13: Entropy production in the brain at different levels of coarse-graining. (*a*) Entropy production at rest and during seven cognitive tasks as a function of the number of clusters k used in hierarchical clustering. The raw entropy production (Eq. 9.9) is divided by the fMRI repetition time $\Delta t = 0.72$ s to compute an entropy production rate, and error bars reflect two standard deviations estimated using trajectory bootstrapping. (*b*) Slope of the relationship between entropy production and physical response rate across tasks for different numbers of clusters k. Error bars represent one-standard-deviation confidence intervals of the slope and asterisks indicate the significance of the correlation between entropy production and response rate. (*c*) Difference between the entropy production during high-load and that during low-load conditions of the working memory task as a function of the number of cluster k. Error bars represent two standard deviations estimated using trajectory bootstrapping, and the entropy production difference is significant across all values of k (one-sided *t*-test, $p < 0.001$).

9.8.10 *Robustness to head motion and signal variance*

In the main text, we showed that the brain's entropy production is significantly correlated with the frequency of physical responses (Fig. 9.4b) and increases during periods of cognitive exertion (Fig. 9.4c). Here, we show that the effects of physical and cognitive effort on entropy production cannot be explained by head movement within the scanner (a common confound in fMRI studies (*231*)) nor variance in the neural time-series. To quantify head movement, for each time point in every time-series, we compute the spatial standard deviation of the difference between the current image and the previous image. This quantity, known as DVARS, is a common measure of head movement in fMRI data (*539*). Importantly, we find that entropy production is not significantly correlated with the average DVARS within each task (Fig. 9.14a), thereby demonstrating that the relationship between entropy production and physical response rate is not simply due to the confound of subject head movement within the scanner. Additionally, we find that entropy production is not significantly correlated with the variance of the neural data within each task (Fig. 9.14b). This final result establishes that our entropy production estimates are not simply driven by variations in the amount of noise in the neural data across different tasks.

9.8.11 *Data processing*

The resting, emotional processing, working memory, social inference, language processing, relational matching, gambling, and motor execution fMRI scans are from the S1200 Human Connectome Project release (*52*, *678*). Brains were normalized to fslr32k via the

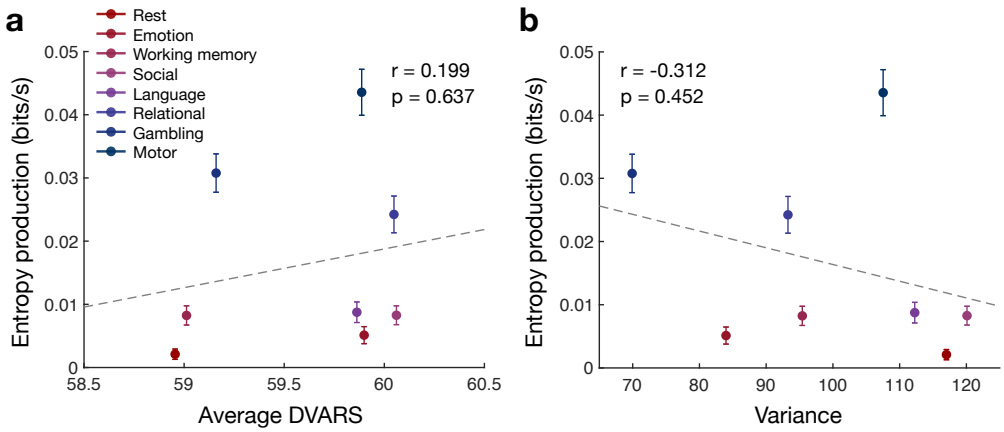


Figure 9.14: Entropy production in the brain cannot be explained by head movement nor signal variance. Entropy production versus the average DVARS (*a*) and the variance of the neural time-series (*b*) at rest and during seven cognitive tasks. Across both panels, entropy productions are estimated using hierarchical clustering with $k = 8$ clusters and are divided by the fMRI repetition time $\Delta t = 0.72$ s to compute entropy production rates. Error bars reflect two standard deviations estimated using trajectory bootstrapping.

MSM-AII registration with 100 regions (*581*). CompCor, with five principal components from the ventricles and white matter masks, was used to regress out nuisance signals from the time series. In addition, the 12 detrended motion estimates provided by the Human Connectome Project were regressed out from the regional time series. The mean global signal was removed and then time series were band-pass filtered from 0.009 to 0.08 Hz. Then, frames with greater than 0.2 mm frame-wise displacement or a derivative root mean square (DVARS) above 75 were removed as outliers. We filtered out sessions composed of greater than 50 percent outlier frames, and we only analyzed data from subjects that had all scans remaining after this filtering, leaving 590 individuals. The processing pipeline used here has previously been suggested to be ideal for removing false relations between neural dynamics and behavior (*612*). Finally, for each subject and each scan, we only analyze the first 176 vectors in the time-series, corresponding to the length of the shortest task (emotional processing); this truncation controls for the possibility of data size affecting comparisons across tasks.

BIBLIOGRAPHY

1. L. F. Abbott, P Dayan, *Theoretical Neuroscience* (MIT Press, 2001).
2. S Achard, R Salvador, B Whitcher, J Suckling, E Bullmore, *J Neurosci* **26**, 63–72 (2006).
3. D. H. Ackley, G. E. Hinton, T. J. Sejnowski, *Cog. Sci.* **9**, 147–169 (1985).
4. A. R. Adamantidis, F. Zhang, A. M. Aravanis, K. Deisseroth, L. De Lecea, *Nature* **450**, 420 (2007).
5. L. A. Adamic, N. Glance, presented at the Proceedings of the 3rd international workshop on Link discovery, pp. 36–43.
6. L. A. Adamic, R. M. Lukose, A. R. Puniyani, B. A. Huberman, *Phys. Rev. E* **64**, 046135 (2001).
7. D. R. Addis, A. T. Wong, D. L. Schacter, *Psychol. Sci.* **19**, 33–41 (2008).
8. J. S. Adelman, G. D. Brown, J. F. Quesada, *Psychol. Sci.* **17**, 814–823 (2006).
9. G. K. Aguirre, *Hastings Cent. Rep.* **44**, S8–S18 (2014).
10. C Aicher, A. Z. Jacobs, A Clauset, *Journal of Complex Networks* **3**, 221–248 (2015).
11. R. Albert, A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
12. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **401**, 130 (1999).
13. R. Albert, H. Jeong, A.-L. Barabási, *Nature* **406**, 378 (2000).
14. A. P. Alivisatos *et al.*, *ACS Nano* **7**, 1850–1866 (2013).
15. G. T. Altmann, Y. Kamide, *Cognition* **73**, 247–264 (1999).
16. S.-i. Amari, H. Nakahara, S. Wu, Y. Sakai, *Neural Comput.* **15**, 127–142 (2003).
17. V. E. Amassian, P. J. Maccabee, R. Q. Cracco, J. B. Cracco, A. P. Rudell, L. Eberle, *Brain Res.* **605**, 317–321 (1993).
18. M Amit, Y Shmerler, E Eisenberg, M Abraham, N Shnerb, *Fractals* **2**, 7–13 (1994).
19. K Amunts, K Zilles, *Neuron* **88**, 1086–1107 (2015).
20. S. S. Andersen, A. D. Jackson, T. Heimburg, *Prog. Neurobiol.* **88**, 104–113 (2009).
21. J. R. Anderson, L. M. Reder, *J. Exp. Psychol.* **128**, 186 (1999).
22. J. R. Anderson, *Cogn. Psychol.* **6**, 451–474 (1974).
23. J. Y. Angela, J. D. Cohen, presented at the Advances in Neural Information Processing Systems, pp. 1873–1880.
24. P. Arena, L. Patané, P. S. Termini, presented at the Neural Networks (IJCNN), The 2010 International Joint Conference on, pp. 1–8.

25. A. Arenas, A. Diaz-Guilera, C. J. Pérez-Vicente, *Phys. Rev. Lett.* **96**, 114102 (2006).
26. A. Arenas, A. Fernandez, S. Gomez, *New J. Phys.* **10**, 053039 (2008).
27. Aristotle, *Metaphysics*, vol. VII.7, 1072b13–30.
28. A Arnatkeviciute, B. D. Fulcher, R Pocock, A Fornito, *PLoS Comput Biol* **14**, e1005989 (2018).
29. R. N. Aslin, E. L. Newport, *Curr. Dir. Psychol. Sci.* **21**, 170–176 (2012).
30. R. N. Aslin, E. L. Newport, *Lang. Learn.* **64**, 86–105 (2014).
31. C. M. Atance, D. K. O'Neill, *Trends Cogn. Sci.* **5**, 533–539 (2001).
32. F. Attneave, *Psychol. Rev.* **61**, 183 (1954).
33. E. Aurell, M. Ekeberg, *Phys. Rev. Lett.* **108**, 090201 (2012).
34. J. Austen, *Pride and prejudice* (Broadview Press, 2001).
35. B. B. Avants, N. J. Tustison, G Song, P. A. Cook, A Klein, J. C. Gee, *Neuroimage* **54**, 2033–2044 (2011).
36. A Avena-Koenigsberger, B Misic, O Sporns, *Nat Rev Neurosci* **19**, 17–33 (2017).
37. J. Avery, M. N. Jones, presented at the Proceedings of the 40th Annual Meeting of the 40th Annual Meeting of the Cognitive Science Society.
38. A Azulay, E Itskovits, A Zaslaver, *PLoS Comput Biol* **12**, e1005021 (2016).
39. R. H. Baayen, D. J. Davidson, D. M. Bates, *J. Mem. Lang.* **59**, 390–412 (2008).
40. J. S. Bach, *The Well-Tempered Clavier, Book I, No. 13*, 1722.
41. A. D. Baddeley, G. Hitch, *Mem. Cogn.* **21**, 146–155 (1993).
42. J. P. Bagrow, D. Wang, A.-L. Barabasi, *PloS one* **6**, e17680 (2011).
43. D. L. Bailey, M. N. Maisey, D. W. Townsend, P. E. Valk, *Positron emission tomography* (Springer, 2005).
44. D. Baldwin, A. Andersson, J. Saffran, M. Meyer, *Cognition* **106**, 1382–1407 (2008).
45. D. A. Balota, M. J. Cortese, S. D. Sergent-Marshall, D. H. Spieler, M. J. Yap, *J. Exp. Psychol.* **133**, 283 (2004).
46. J Bancaud, J Talairach, *Rev. Otoneuroophtalmol.* **45**, 315—328 (1973).
47. M. J. Banissy, R. Kanai, V. Walsh, G. Rees, *Neuroimage* **62**, 2034–2039 (2012).
48. Y. Bar-Hillel, R. Carnap, *Br. J. Philos. Sci.* **4**, 147–157 (1953).
49. A.-L. Barabási, *Nature* **435**, 207–211 (2005).
50. A.-L. Barabási, R. Albert, *Science* **286**, 509–512 (1999).
51. A.-L. Barabâsi, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek, *Physica A* **311**, 590–614 (2002).
52. D. M. Barch, G. C. Burgess, M. P. Harms, S. E. Petersen, B. L. Schlaggar, M. Corbetta, M. F. Glasser, S. Curtiss, S. Dixit, C. Feldt, *et al.*, *Neuroimage* **80**, 169–189 (2013).

53. H. B. Barlow (1961).

54. H. Barlow, *Vision Res.* **30**, 1561–1571 (1990).

55. A. Baronchelli, R. Ferrer-i Cancho, R. Pastor-Satorras, N. Chater, M. H. Christiansen, *Trends Cogn. Sci.* **17**, 348–360 (2013).

56. P. J. Basser, S Pajevic, C Pierpaoli, J Duda, A Aldroubi, *Magn Reson Med* **44**, 625–632 (2000).

57. D. S. Bassett, E. T. Bullmore, *Neuroscientist* **Sep 21**, 1073858416667720 (2016).

58. D. S. Bassett, E Bullmore, *Neuroscientist* **12**, 512–523 (2006).

59. D. S. Bassett, M. S. Gazzaniga, *Trends Cogn Sci* **15**, 200–209 (2011).

60. D. S. Bassett, O Sporns, *Nat Neurosci* **20**, 353–364 (2017).

61. D. S. Bassett, D. L. Greenfield, A Meyer-Lindenberg, D. R. Weinberger, S. W. Moore, E. T. Bullmore, *PLoS Comput. Biol.* **6**, e1000748 (2010).

62. D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, S. T. Grafton, *Chaos* **24**, 013112 (2014).

63. D. S. Bassett, P. Zurn, J. I. Gold, *Nat. Rev. Neurosci.* **19**, 566–578 (2018).

64. V. Batagelj, A. Mrvar, <*vlado.fmf.uni-lj.si/pub/networks/data/* > (2006).

65. D. Bates, M. Mächler, B. Bolker, S. Walker, *et al.*, *J. Stat. Softw.* **67** (2015).

66. D Battaglia, A Witt, F Wolf, T Geisel, *PLoS Comput Biol* **8**, e1002438 (2012).

67. C. Battle, C. P. Broedersz, N. Fakhri, V. F. Geyer, J. Howard, C. F. Schmidt, F. C. MacKintosh, *Science* **352**, 604–607 (2016).

68. G. L. Baum *et al.*, *Curr Biol* **27**, 1561–1572.e8 (2017).

69. A. Bechara, A. R. Damasio, H. Damasio, S. W. Anderson, *Cognition* **50**, 7–15 (1994).

70. A. Beck, *Centralbl Physiol* **4**, 572–573 (1890).

71. N. Beckage, L. Smith, T. Hills, *PloS One* **6**, e19348 (2011).

72. C. O. Becker, D. S. Bassett, V. M. Preciado, *J. Neural Eng.* **15**, 066016 (2018).

73. L. v. Beethoven, *Piano Sonata No. 23*, 1807.

74. J. M. Beggs, D. Plenz, *Journal of Neuroscience* **23**, 11167–11177, ISSN: 0270-6474 (2003).

75. T. E. Behrens, H Johansen-Berg, *Philos Trans R Soc Lond B Biol Sci* **360**, 903–911 (2005).

76. B Bentley, R Branicky, C. L. Barnes, Y. L. Chew, E Yemini, E. T. Bullmore, P. E. Vertes, W. R. Schafer, *PLoS Comput Biol* **12**, e1005283 (2016).

77. A. Berényi, M. Belluscio, D. Mao, G. Buzsáki, *Science* **337**, 735–737 (2012).

78. N. Bertschinger, T. Natschläger, *Neural Computation* **16**, 1413–1436 (2004).

79. L. M. Bettencourt, G. J. Stephens, M. I. Ham, G. W. Gross, *Phys Rev E* **75**, 021915 (2007).

80. R. F. Betzel, D. S. Bassett, *J R Soc Interface* **14**, 20170623 (2017).

81. R. F. Betzel, D. S. Bassett, *Neuroimage* **160**, 73–83 (2017).

82. R. F. Betzel, D. S. Bassett, *Proc Natl Acad Sci U S A* **115**, E4880–E4889 (2018).

83. R. F. Betzel, J. D. Medaglia, D. S. Bassett, *Nature Communications* **9**, 346 (2018).

84. R. F. Betzel, A. Avena-Koenigsberger, J. Goñi, Y. He, M. A. de Reus, A. Griffa, P. E. Vértes, B. Mišic, J.-P. Thiran, P. Hagmann, *et al.*, *Neuroimage* **124**, 1054–1064 (2016).

85. R. F. Betzel, S. Gu, J. D. Medaglia, F. Pasqualetti, D. S. Bassett, *Sci. Rep.* **6**, 30770 (2016).

86. S. F. Beul, S Grant, C. C. Hilgetag, *Brain Struct Funct* **220**, 3167–3184 (2015).

87. S. F. Beul, H Barbas, C. C. Hilgetag, *Sci Rep* **7**, 43176 (2017).

88. R. L. Beurle, *Phil. Trans. R. Soc. Lond. B* **240**, 55–94 (1956).

89. W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, A. M. Walczak, *Proc. Natl. Acd. Sci.* **109**, 4786–4791 (2012).

90. G Bianconi, C Rahmede, Z Wu, *Phys Rev E Stat Nonlin Soft Matter Phys* **92**, 022815 (2015).

91. W. Blake, *Songs of Innocence and of Experience* (Princeton University Press, 1998), vol. 5.

92. L. Blume, *GEB* **5**, 387–424 (1993).

93. M. L. Boas, *Mathematical methods in the physical sciences* (Wiley, 2006).

94. P. Boldi, B. Codenotti, M. Santini, S. Vigna, *Software Pract. Exper.* **34**, 711–726 (2004).

95. E. Bolthausen, *Comm. Math. Phys.* **325**, 333–366 (2014).

96. J. Borge-Holthoefer, A. Arenas, *Entropy* **12**, 1264–1302 (2010).

97. J. Borge-Holthoefer, N. Perra, B. Gonçalves, S. González-Bailón, A. Arenas, Y. Moreno, A. Vespignani, *Sci. Adv.* **2**, e1501158 (2016).

98. E Boto *et al.*, *Nature* **555**, 657–661 (2018).

99. W. A. Bousfield, *J. Gen. Psychol.* **49**, 229–240 (1953).

100. E. S. Boyden, F Zhang, E Bamberg, G Nagel, K Deisseroth, *Nat Neurosci* **8**, 1263–1268 (2005).

101. T. F. Brady, A. Oliva, *Psychol. Sci.* **19**, 678–685 (2008).

102. J. Brahms, *Ballades, Op. 10, No. 1*, 1854.

103. C. P. Brangwynne, G. H. Koenderink, F. C. MacKintosh, D. A. Weitz, *J. Cell Biol.* **183**, 583–587 (2008).

104. U Braun, A Schaefer, R. F. Betzel, H Tost, A Meyer-Lindenberg, D. S. Bassett, *Neuron* **97**, 14–31 (2018).

105. A. J. Bray, H. Sompolinsky, C. Yu, *J. Phys. C: Solid State Phys.* **19**, 6389–6406 (1986).

106. M Breakspear, *Nat Neurosci* **20**, 340–352 (2017).

107. M. R. Brent, T. A. Cartwright, *Cognition* **61**, 93–125 (1996).

108. P. Broca, *Bulletin et Memoires de la Societe anatomique de Paris* **6**, 330–357 (1861).

109. C. D. Brody, *Neural Comput* **11**, 1537–1551 (1999).

110. C. D. Brody, *Neural Comput* **11**, 1527–1535 (1999).

111. G. J. Brown, M. Cooke, *Comput. Speech Lang.* **8**, 297–336 (1994).

112. P. F. Brown, P. V. Desouza, R. L. Mercer, V. J. D. Pietra, J. C. Lai, *Comput. Linguist.* **18**, 467–479 (1992).

113. S. G. Brush, *Rev. Mod. Phys.* **39**, 883 (1967).

114. J. Bryden, N. Cohen, presented at the From Animals to Animats 8: Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior, pp. 183–192.

115. E Bullmore, O Sporns, *Nat Rev Neurosci* **10**, 186–198 (2009).

116. E. Bullmore, O. Sporns, *Nat. Rev. Neurosci.* **13**, 336–349 (2012).

117. Z. Burda, J. Duda, J.-M. Luck, B. Waclaw, *Phys. Rev. Lett.* **102**, 160602 (2009).

118. C. T. Butts, *Science* **325**, 414–416 (2009).

119. J. Cabral, E. Hugues, O. Sporns, G. Deco, *Neuroimage* **57**, 130–139 (2011).

120. G. Caldarelli, M. Marsili, Y.-C. Zhang, *EPL* **40**, 479 (1997).

121. R. F. I. Cancho, R. V. Solé, *Proc. R. Soc. Lond., B, Biol. Sci.* **268**, 2261–2265 (2001).

122. J. Candia, M. C. González, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabási, *J. Phys. A* **41**, 224015 (2008).

123. M. T. Carlson, M. Sonderegger, M. Bane, *J. Mem. Lang.* **75**, 159–180 (2014).

124. S. Carnot, *Reflexions sur la puissance motrice du feu* (Bachelier, Paris, France, 1824).

125. S. Cash, R. Yuste, *Neuron* **22**, 383–394 (1999).

126. C. Castellano, S. Fortunato, V. Loreto, *Rev. Mod. Phys.* **81**, 591 (2009).

127. C. Castillo, K. Chellapilla, L. Denoyer, presented at the Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web.

128. M. Catanzaro, R. Pastor-Satorras, *Eur. Phys. J. B* **44**, 241–248 (2005).

129. R. Caton, *J. Nerv. Ment. Dis.* **2**, 610 (1875).

130. A. Cavagna, I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone, A. M. Walczak, *Phys. Rev. E* **89**, 042707 (2014).

131. O. Celma, in *Music recommendation and discovery* (Springer, 2010), pp. 43–85.

132. M. Cervantes, *Don Quixote* (LBA, 2018).

133. K. Y. Chan, M. S. Vitevitch, *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 1934 (2009).

134. K. Y. Chan, M. S. Vitevitch, *Cogn. Sci.* **34**, 685–697 (2010).

135. D. Chandler, *Introduction to Modern Statistical Mechanics* (Oxford University Press, 1987).

136. R Chaudhuri, I Fiete, *Nat Neurosci* **19**, 394–403 (2016).

137. R Chaudhuri, K Knoblauch, M. A. Gariel, H Kennedy, X. J. Wang, *Neuron* **88**, 419–431 (2015).

138. P Chauvel, J Vignal, A Biraben, J Badier, J Scarabin, in *Multimethodological Assessment of the Epileptic Forms*, ed. by G Pawlik, H Stefan (Springer Verlag, 1996), pp. 80–108.

139. W. Chen, Y. Wang, S. Yang, presented at the SIGKDD, pp. 199–208.

140. W. Chen, C. Wang, Y. Wang, presented at the SIGKDD, pp. 1029–1038.

141. D. R. Chialvo, *Nat. Phys.* **6**, 744 (2010).

142. S. Chiken, A. Nambu, *Front. Syst. Neurosci.* **8**, 33 (2014).

143. S. Ching, M. Y. Liberman, J. J. Chemali, M. B. Westover, J. D. Kenny, K. Solt, P. L. Purdon, E. N. Brown, *Anesthesiology* **119**, 848–860 (2013).

144. F. Chopin, *Nocturnes, Op. 9, No. 2*, 1832.

145. N. Chopra, M. W. Spong, *IEEE Trans. Automat. Contr.* **54**, 353–357 (2009).

146. R Ciric *et al.*, *Neuroimage* **154**, 174–187 (2017).

147. A. Cleeremans, J. L. McClelland, *J. Exp. Psychol. Gen.* **120**, 235 (1991).

148. F. Coghi, J. Morand, H. Touchette, *Phys. Rev. E* **99**, 022137 (2019).

149. J. E. Cohen, *Behav. Sci.* **7**, 137–163 (1962).

150. J. D. Cohen, S. M. McClure, J. Y. Angela, *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **362**, 933–942 (2007).

151. M. R. Cohen, A Kohn, *Nat Neurosci* **14**, 811–819 (2011).

152. J. S. Coleman *et al.*, *Introduction to mathematical sociology.* (London Free Press Glencoe, 1964).

153. A. G. Collins, M. J. Frank, *Eur. J. Neurosci.* **35**, 1024–1035 (2012).

154. A. G. Collins, M. J. Frank, *Proc. Natl. Acad. Sci.* 201720963 (2018).

155. A. O. Constantinescu, J. X. O'Reilly, T. E. J. Behrens, *Science* **352**, 1464–1468 (2016).

156. E. J. Cornblath, A. Ashourvan, J. Z. Kim, R. F. Betzel, R. Ciric, G. L. Baum, X. He, K. Ruparel, T. M. Moore, R. C. Gur, *et al.*, *Communications Biology, in press*.

157. S. P. Cornelius, W. L. Kath, A. E. Motter, *Nat. Commun.* **4**, 1942 (2013).

158. J.-M. Coron, *Control and nonlinearity* (American Mathematical Soc., 2007).

159. L. d. F. Costa, F. A. Rodrigues, G Travieso, P. R. Villas Boas, *Advances In Physics* **56**, 167–242 (2006).

160. J. P. Coughlin, R. H. Baran, *Neural computation in hopfield networks and boltzmann machines* (University of Delaware Press, 1995).

161. T. M. Cover, J. A. Thomas, *Elements of information theory* (John Wiley & Sons, 2012).

162. M. Crandall, P. Rabinowitz, *Journal of Functional Analysis* **8**, 321–340 (1971).

163. R. Crane, D. Sornette, *Proc. Natl. Acd. Sci.* **105**, 15649–15653 (2008).

164. J. Crinion, J. Ashburner, A. Leff, M. Brett, C. Price, K. Friston, *Neuroimage* **37**, 866–875 (2007).

165. J. P. Cunningham, M. Y. Byron, *Nat. Neurosci.* **17**, 1500 (2014).

166. M. R. D'Orsogna, M. Perc, *Phys. Life Rev.* **12**, 1–21 (2015).

167. J. Dall, M. Christensen, *Physical Review E* **66**, 016121 (2002).

168. O. David, K. J. Friston, *NeuroImage* **20**, 1743–1755 (2003).

169. O. David, D. Cosmelli, K. J. Friston, *Neuroimage* **21**, 659–673 (2004).

170. P. Dayan, *Neural Comput.* **5**, 613–624 (1993).

171. M De Domenico, C Granell, M. A. Porter, A Arenas, **12**, 901–906 (2016).

172. W. De Nooy, A. Mrvar, V. Batagelj, *Exploratory social network analysis with Pajek* (Cambridge University Press, 2011), vol. 27.

173. A. De, I. Valera, N. Ganguly, S. Bhattacharya, M. Gomez-Rodriguez, *arXiv preprint* (2015).

174. T. de Camp Wilson, R. E. Nisbett, *Soc. Psychol.* 118–131 (1978).

175. D. J. de Solla Price, *Science* **149**, 510—515 (1965).

176. T. W. Deacon, *The symbolic species: The co-evolution of language and the brain* (WW Norton & Company, 1998).

177. A. Decelle, F. Krzakala, C. Moore, L. Zdeborová, *Phys. Rev. E* **84**, 066106 (2011).

178. J. B. Dechery, J. N. MacLean, *J Neurophysiol* **118**, 1914–1925 (2017).

179. G. Deco, G. Tononi, M. Boly, M. L. Kringelbach, *Nat. Rev. Neurosci.* **16**, 430 (2015).

180. S. Dehaene, F. Meyniel, C. Wacongne, L. Wang, C. Pallier, *Neuron* **88**, 2–19 (2015).

181. K. Deisseroth, *Nat. Methods* **8**, 26 (2011).

182. A. Delgado-Bonal, J. Martín-Torres, *Sci. Rep.* **6**, 36038 (2016).

183. L. Demetrius, T. Manke, *Physica A* **346**, 682–696 (2005).

184. M. Derex, R. Boyd, *Nat. Commun.* **6**, 1–7 (2015).

185. F. Deschâtres, D. Sornette, *Phys. Rev. E* **72**, 016112 (2005).

186. R Dias, T. Robbins, A. Roberts, *Nature* **380**, 69 (1996).

187. E Dieterich, J Camunas-Soler, M Ribezzi-Crivellari, U Seifert, F Ritort, *Nat. Phys.* **11**, 971–977 (2015).

188. A. Dix, *Human-computer interaction* (Springer, 2009).

189. P. S. Dodds, R. Muhamad, D. J. Watts, *Science* **301**, 827–829 (2003).

190. P. Domingos, M. Richardson, presented at the KDD, pp. 57–66.

191. B. P. Dore, C Scholz, E. C. Baek, J. O. Garcia, M. B. O'Donnell, D. S. Bassett, J. M. Vettel, E. B. Falk, *Cereb Cortex*, Aug 28 (2018).

192. S. N. Dorogovtsev, J. F. F. Mendes, *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **268**, 2603–2606 (2001).

193. F. I. Dretske, *Knowledge and the Flow of Information*.

194. F. Du, X.-H. Zhu, Y. Zhang, M. Friedman, N. Zhang, K. Uğurbil, W. Chen, *Proc. Natl. Acad. Sci.* **105**, 6409–6414 (2008).

195. H. Dubossarsky, S. De Deyne, T. T. Hills, *Dev. Psychol.* **53**, 1560 (2017).

196. J. Duchi, S. Shalev-Shwartz, Y. Singer, T. Chandra, *ICML*, 272–279 (2008).

197. M. S. e Silva, M. Depken, B. Stuhrmann, M. Korsten, F. C. MacKintosh, G. H. Koenderink, *Proc. Natl. Acad. Sci.* **108**, 9408–9413 (2011).

198. J.-P. Eckmann, E. Moses, D. Sergi, *Proc. Natl. Acd. Sci.* **101**, 14333–14337 (2004).

199. F. A. Edwards, A. Konnerth, B. Sakmann, T. Takahashi, *Pflügers Archiv* **414**, 600–612 (1989).

200. M. Egmont-Petersen, D. de Ridder, H. Handels, *Pattern Recognit.* **35**, 2279–2301 (2002).

201. D. A. Egolf, *Science* **287**, 101–104 (2000).

202. T. Engelthaler, T. T. Hills, *Cogn. Sci.* **41**, 120–140 (2017).

203. M Ercsey-Ravasz, N. T. Markov, C Lamy, D. C. Van Essen, K Knoblauch, Z Toroczkai, H Kennedy, *Neuron* **80**, 184–197 (2013).

204. P. Erdös, A. Rényi, *Publ. Math. Debrecen* **6**, 290–297 (1959).

205. P. Erdös, A. Rényi, *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–60 (1960).

206. M. Erecińska, I. A. Silver, *J. Cereb. Blood Flow Metab.* **9**, 2–19 (1989).

207. K. A. Eriksen, I. Simonsen, S. Maslov, K. Sneppen, *Phys. Rev. Lett.* **90**, 148701 (2003).

208. M. Esposito, *Phys. Rev. E* **85**, 041125 (2012).

209. E. Estrada, N. Hatano, *Phys. Rev. E* **77**, 036111 (2008).

210. E. Estrada, N. Hatano, M. Benzi, *Phys. Rep.* **514**, 89–119 (2012).

211. D. J. Evans, E. G. D. Cohen, G. P. Morriss, *Phys. Rev. Lett.* **71**, 2401 (1993).

212. E. B. Falk, D. S. Bassett, *Trends Cogn Sci* **21**, 674–690 (2017).

213. R. Falk, C. Konold, *Psychol. Rev.* **104**, 301 (1997).

214. J. E. Ferrell, E. M. Machleder, *Science* **280**, 895–898 (1998).

215. A. M. Ferrenberg, R. H. Swendsen, *Phys. Rev. Lett.* **61**, 2635 (1988).

216. M. Fielder, V. Ptak, *Czech. Math. J.* **22**, 382–400 (1962).

217. J. Fiser, R. N. Aslin, *J. Exp. Psychol.* **28**, 458 (2002).

218. R. FitzHugh, *Biophys. J.* **1**, 445–466 (1961).

219. S. M. Fleming, R. S. Weil, Z. Nagy, R. J. Dolan, G. Rees, *Science* **329**, 1541–1543 (2010).

220. P. Flourens, *Recherches expérimentales sur les propriétés et les fonctions du système nerveux dans les animaux vertébrés* (Ballière, 1842).

221. K. I. Forster, S. M. Chambers, *J. Verbal Learning Verbal Behav.* **12**, 627–635 (1973).

222. C. M. Fortuin, P. W. Kasteleyn, J. Ginibre, *Comm. in Math. Phys.* **22**, 89–103 (1971).

223. S Fortunato, D Hric, *Physics Reports* **659**, 1–44 (2016).

224. S. Fortunato, *Physics reports* **486**, 75–174 (2010).

225. J. G. Foster, D. V. Foster, P. Grassberger, M. Paczuski, *Proc. Natl. Acad. Sci.* **107**, 10815–10820 (2010).

226. M. D. Fox, M. Corbetta, A. Z. Snyder, J. L. Vincent, M. E. Raichle, *Proc. Natl. Acad. Sci.* **103**, 10046–10051 (2006).

227. A. D. Friederici, *Trends Cogn. Sci.* **9**, 481–488 (2005).

228. P Fries, *Neuron* **88**, 220–235 (2015).

229. P. Fries, *Trends Cogn. Sci.* **9**, 474–480 (2005).

230. K. Friston, C. Frith, P. Liddle, R. Frackowiak, *J. Cereb. Blood Flow Metab.* **13**, 5–14 (1993).

231. K. J. Friston, S. Williams, R. Howard, R. S. Frackowiak, R. Turner, *Magn. Reson. Med.* **35**, 346–355 (1996).

232. K. Friston, J. Kilner, L. Harrison, *J. Physiol. Paris* **100**, 70–87 (2006).

233. K. Friston, S. Samothrakis, R. Montague, *Biological cybernetics* **106**, 523–541 (2012).

234. D. Fudenberg, D. Levine, *The Theory of Learning in Games* (MIT Press, Cambridge, MA, 1998).

235. J. M. Fuster, G. E. Alexander, *Science* **173**, 652–654 (1971).

236. S. Galam, *Physica A* **230**, 174–188 (1996).

237. S. Galam, *Int. J. Mod. Phys. C* **19**, 409–440 (2008).

238. S. Galam, S. Moscovici, *Eur. J. Soc. Psychol.* **21**, 49–74 (1991).

239. S. Galam, Y. Gefen, Y. Shapir, *J. Math. Sociol.* **9**, 1–13 (1982).

240. S. Galam *et al.*, *Physica A* **238**, 66–80 (1997).

241. E. Ganmor, R. Segev, E. Schneidman, *Proc. Natl. Acd. Sci.* **108**, 9679–9684 (2011).

242. M. M. Garvert, R. J. Dolan, T. E. Behrens, *Elife* **6** (2017).

243. M. S. George, S. H. Lisanby, H. A. Sackeim, *Archives of General Psychiatry* **56**, 300–311 (1999).

244. A. Georges, J. S. Yedidia, *Journal of Physics A: Mathematical and General* **24** (1991).

245. D. L. Gerlough, A. Schuhl, *Use of Poisson distribution in highway traffic* (Eno Foundation for Highway Traffic Control, 1955).

246. S. J. Gershman, *J. Neurosci.* **38**, 7193–7200 (2018).

247. S. J. Gershman, C. D. Moore, M. T. Todd, K. A. Norman, P. B. Sederberg, *Neural Comput.* **24**, 1553–1568 (2012).

248. S. J. Gershman, E. J. Horvitz, J. B. Tenenbaum, *Science* **349**, 273–278 (2015).

249. W. R. Gilks, S. Richardson, D. Spiegelhalter, *Markov chain Monte Carlo in practice* (CRC press, 1995).

250. M. Girvan, M. E. Newman, *Proc. Natl. Acad. Sci.* **99**, 7821–7826 (2002).

251. C Giusti, E Pastalkova, C Curto, V Itskov, *Proc Natl Acad Sci U S A* **112**, 13455–13460 (2015).

252. C Giusti, R Ghrist, D. S. Bassett, *J Comput Neurosci* **41**, 1–14 (2016).

253. D. Gleich, L. Zhukov, P. Berkhin, presented at the Yahoo! Research Technical Report YRL-2004-038, vol. 13, p. 22.

254. P. M. Gleiser, L. Danon, *Adv. Complex Syst.* **6**, 565–573 (2003).

255. F. Gnesotto, F. Mura, J. Gladrow, C. Broedersz, *Rep. Prog. Phys.* **81**, 066601 (2018).

256. G. V. Goddard, *Nature* **214**, 1020–1021 (1967).

257. G. V. Goddard, D. C. McIntyre, C. K. Leech, *Exp. Neurol.* **25**, 295–330 (1969).

258. K.-I. Goh, B. Kahng, D. Kim, *Phys. Rev. Lett.* **87**, 278701 (2001).

259. P. S. Goldman, H. E. Rosvold, *Exp. Neurol.* **27**, 291–304 (1970).

260. R. Goldstein, M. S. Vitevitch, *Front. Psychol.* **5**, 1307 (2014).

261. C. Golgi, *Sulla fina anatomia degli organi centrali del sistema nervoso* (S. Calderini, 1885).

262. R. G. Golledge, in *The Colonization of Unfamiliar Landscapes* (Routledge, 2003), pp. 49–54.

263. J. Gómez-Gardeñes, V. Latora, *Phys. Rev. E* **78**, 065102 (2008).

264. A. Gomez-Marin, J. M. Parrondo, C. Van den Broeck, *Phys. Rev. E* **78**, 011107 (2008).

265. M. Gomez-Rodriguez, S Bernhard, *ICML* (2012).

266. R. L. Gómez, *Psychol. Sci.* **13**, 431–436 (2002).

267. R. L. Gomez, L. Gerken, *Cognition* **70**, 109–135 (1999).

268. G. Gong, Y. He, L. Concha, C. Lebel, D. W. Gross, A. C. Evans, C. Beaulieu, *Cereb. cortex* **19**, 524–536 (2008).

269. G. Gong, P. Rosa-Neto, F. Carbonell, Z. J. Chen, Y. He, A. C. Evans, *J. Neurosci.* **29**, 15684–15693 (2009).

270. J. Goñi, M. P. van den Heuvel, A. Avena-Koenigsberger, N. V. de Mendizabal, R. F. Betzel, A. Griffa, P. Hagmann, B. Corominas-Murtra, J.-P. Thiran, O. Sporns, *Proceedings of the National Academy of Sciences* **111**, 833–838 (2014).

271. M. C. Gonzalez, C. A. Hidalgo, A.-L. Barabasi, *Nature* **453**, 779–782 (2008).

272. A. Gopnik, H. M. Wellman, *Psychol. Bull.* **138**, 1085 (2012).

273. M Gosak, R Markovic, J Dolensek, M Slak Rupnik, M Marhl, A Stozer, M Perc, *Phys Life Rev* **24**, 118–135 (2018).

274. A. Goyal, F. Bonchi, L. V. Lakshmanan, *VLDB Endowment* **5**, 73–84 (2011).

275. S. Goyal, H. Heidari, M. Kearns, *GEB* (2014).

276. D. J. Green, R. Gillette, *Brain Res.* **245**, 198–200 (1982).

277. R. L. Gregory, *Phil. Trans. R. Soc. Lond. B* **290**, 181–197 (1980).

278. B. F. Grewe, D. Langer, H. Kasper, B. M. Kampa, F. Helmchen, *Nat. Methods* **7**, 399 (2010).

279. R. Griffiths, C. Hurst, S. Sherman, *J. Math. Phys.* **11**, 790 (1970).

280. R. B. Griffiths, *J. Math. Phys.* **8**, 478–483 (1967).

281. G. Grimmett, D. Stirzaker, *Probability and random processes* (Oxford university press, 2001).

282. L. Grosenick, J. H. Marshel, K. Deisseroth, *Neuron* **86**, 106–139 (2015).

283. T. Gross, B. Blasius, *J R Soc Interface* **5**, 259–271 (2008).

284. S Gu *et al.*, *Nat Commun* **6**, 8414 (2015).

285. S. Gu, R. F. Betzel, M. G. Mattar, M. Cieslak, P. R. Delio, S. T. Grafton, F. Pasqualetti, D. S. Bassett, *Neuroimage* **148**, 305–317 (2017).

286. R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, A. Arenas, *Phys. Rev. E* **68**, 065103 (2003).

287. L. A. Gunaydin, O. Yizhar, A. Berndt, V. S. Sohal, K. Deisseroth, P. Hegemann, *Nature neuroscience* **13**, 387 (2010).

288. L. F. Haas, *J. Neurol. Neurosurg. Psychiatry* **74**, 9–9 (2003).

289. A Hackett, s Melnik, J. P. Gleeson, *Phys. Rev. E* **83**, 056107 (2011).

290. F. A. Haight, *Handbook of the Poisson distribution* (Wiley, New York, 1967).

291. C. Haldeman, J. M. Beggs, *Phys. Rev. Lett.* **94**, 058101 (5 2005).

292. M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, O. V. Lounasmaa, *Rev. Mod. Phys.* **65**, 413 (1993).

293. G. E. Hardingham, P Pruunsild, M. E. Greenberg, H Bading, *Nat Rev Neurosci* **19**, 9–15 (2018).

294. R Hari, R Salmelin, *Neuroimage* **61**, 386–396 (2012).

295. J. J. Harris, R. Jolivet, D. Attwell, *Neuron* **75**, 762–777 (2012).

296. C. A. Hartley, B. Fischl, E. A. Phelps, *Cereb. Cortex* **21**, 1954–1962 (2011).

297. A. F. Hayes, *Statistical methods for communication science* (Routledge, 2009).

298. G. Haynes, H Hermes, *SIAM J. Control* **8**, 450–460 (1970).

299. Y. He, J. Wang, L. Wang, Z. J. Chen, C. Yan, H. Yang, H. Tang, C. Zhu, Q. Gong, Y. Zang, *et al.*, *PloS one* **4**, e5226 (2009).

300. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, B. Scholkopf, *IEEE Intell. Syst.* **13**, 18–28 (1998).

301. D Hebb, *The Organization of Behavior* (Wiley, 1949).

302. C. N. Heck *et al.*, *Epilepsia* **55**, 432–441 (2014).

303. D. Helbing, D. Brockmann, T. Chadefaux, K. Donnay, U. Blanke, O. Woolley-Meza, M. Moussaid, A. Johansson, J. Krause, S. Schutte, *et al.*, *J. Stat. Phys.* **158**, 735–781 (2015).

304. M Helmstaedter, K. L. Briggman, S. C. Turaga, V Jain, H. S. Seung, W Denk, *Nature* **500**, 168–174 (2013).

305. S Henriksen, R Pang, M Wronkiewicz, *Elife* **5**, e12366 (2016).

306. R. Hermann, A. Krener, *IEEE Trans. Automat. Contr.* **22**, 728–740 (1977).

307. A. M. Hermundstad, K. S. Brown, D. S. Bassett, J. M. Carlson, *PLoS Comput Biol* **7**, e1002063 (2011).

308. J. Hertz, A. Krogh, R. G. Palmer, *Introduction to the theory of neural computation.* (Addison-Wesley/Addison Wesley Longman, 1991).

309. M. Hilbert, *Psychol. Bull.* **138**, 211 (2012).

310. C.-C. Hilgetag, G. A. Burns, M. A. O'Neill, J. W. Scannell, M. P. Young, *Phil. Trans. R. Soc. Lon. B* **355**, 91–110 (2000).

311. B. Hille *et al.*, *Ion channels of excitable membranes* (Sinauer Sunderland, MA, 2001), vol. 507.

312. W. D. Hillis, *Daedalus*, 175–189 (1988).

313. T. T. Hills, M. Maouene, J. Maouene, A. Sheya, L. Smith, *Psychol. Sci.* **20**, 729–739 (2009).

314. T. T. Hills, M. N. Jones, P. M. Todd, *Psychol. Rev.* **119**, 431 (2012).

315. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, *IEEE Signal Process. Mag.* **29**, 82–97 (2012).

316. A. L. Hodgkin, A. F. Huxley, *J. Physiol.* **117**, 500–544 (1952).

317. P. Holme, J. Saramäki, *Phys. Rep.* **519**, 97–125 (2012).

318. G. Holmes, *The British journal of ophthalmology* **2**, 353 (1918).

319. A. B. Holt, D Wilson, M Shinn, J Moehlis, T. I. Netoff, *PLoS Comput Biol* **12**, e1005011 (2016).

320. A. B. Holt, T. I. Netoff, *J. Comput. Neurosci.* **37**, 505–521 (2014).

321. C. Honey, O Sporns, L. Cammoun, X. Gigandet, J.-P. Thiran, R. Meuli, P. Hagmann, *Proceedings of the National Academy of Sciences* **106**, 2035–2040 (2009).

322. J. J. Hopfield, *Proc. Nat. Acad. Sci. (USA)* **79**, 2554–2558 (1982).

323. M. W. Howard, M. J. Kahana, *J. Exp. Psychol. Learn. Mem. Cogn.* **25**, 923 (1999).

324. M. W. Howard, M. J. Kahana, *J. Math. Psychol.* **46**, 269–299 (2002).

325. D. Howe, <*foldoc.org* > (1993).

326. J. J. Hox, M. Moerbeek, R. van de Schoot, *Multilevel analysis: Techniques and applications* (Routledge, 2017).

327. J. Hsieh *et al.*, presented at the.

328. K. C. Huang, Y. Meir, N. S. Wingreen, *Proc. Natl. Acad. Sci.* **100**, 12724–12728 (2003).

329. S. A. Huettel, P. B. Mack, G. McCarthy, *Nat. Neurosci.* **5**, 485 (2002).

330. R. Hyman, *J. Exp. Psychol.* **45**, 188 (1953).

331. L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, W. Van den Broeck, *J. Theor. Biol.* **271**, 166–180 (2011).

332. D. Isenberg, *J. of Personality and Social Psychology* **50**, 1141–1151 (1986).

333. A. Isidori, *Nonlinear control systems* (Springer Science & Business Media, 2013).

334. E. Ising, *Zeitschrift für Physik* **31**, 253–258 (1925).

335. F. L. Iudice, F. Garofalo, F. Sorrentino, *Nat. Commun.* **6**, 8349 (2015).

336. M. Jackson, *Thriller*, 1984.

337. G. Jafari, P Pedram, L Hedayatifar, *J. Stat. Mech.: Theory Exp.* (2007).

338. A. K. Jain, M. N. Murty, P. J. Flynn, *ACM Comput. Surv.* **31**, 264–323 (1999).

339. E. T. Jaynes, *Phys. Rev.* **106**, 620 (1957).

340. M. Jazayeri, M. N. Shadlen, *Nat. Neurosci.* **13**, 1020 (2010).

341. J Jeganathan, A Perry, D. S. Bassett, G Roberts, P. B. Mitchell, M Breakspear, *Neuroimage Clin* **19**, 71–81 (2018).

342. Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, K. Xie, presented at the AAAI, vol. 11, pp. 127–132.

343. P. N. Johnson-Laird, *Cogn. Sci.* **4**, 71–115 (1980).

344. E. Jonas, S. Schulz-Hardt, D. Frey, N. Thelen, *J. Pers. Soc. Psychol.* **80**, 557 (2001).

345. M. Jones, T. Hills, P. Todd, *Psychol. Rev.* **122**, 570–574 (2015).

346. M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, *Machine learning* **37**, 183–233 (1999).

347. M. I. Jordan, *Learning in graphical models* (Springer Science & Business Media, 1998), vol. 89.

348. B. Jowett, *The republic of Plato* (Clarendon press, 1888).

349. G. Kachergis, C. Yu, R. M. Shiffrin, *Psychon. Bull. Rev.* **19**, 317–324 (2012).

350. G. Kachergis, C. Yu, R. M. Shiffrin, *Top. Cogn. Sci.* **5**, 200–213 (2013).

351. A. E. Kahn, E. A. Karuza, J. M. Vettel, D. S. Bassett, *Nat. Hum. Behav.* **2**, 936 (2018).

352. T. Kailath, *Linear Systems* (Prentice-Hall, Inc., 1980).

353. M Kaiser, *Trends Cogn Sci* **21**, 703–717 (2017).

354. M. Kaiser, C. C. Hilgetag, *PLOS Comput. Biol.* **2**, e95 (2006).

355. R. Kanai, G. Rees, *Nat. Rev. Neurosci.* **12**, 231 (2011).

356. R. Kanai, T. Feilden, C. Firth, G. Rees, *Curr. Biol.* **21**, 677–680 (2011).

357. J. N. Kapur, *Maximum-entropy models in science and engineering* (John Wiley & Sons, 1989).

358. T. Karagiannis, M. Molle, M. Faloutsos, *IEEE Internet Comput.* **8**, 57–64 (2004).

359. B. Karrer, M. E. Newman, L. Zdeborová, *Phys. Rev. Lett.* **113**, 208702 (2014).

360. E. A. Karuza, S. L. Thompson-Schill, D. S. Bassett, *Trends Cogn. Sci.* **20**, 629–640 (2016).

361. E. A. Karuza, A. E. Kahn, S. L. Thompson-Schill, D. S. Bassett, *Sci. Rep.* **7**, 12733 (2017).

362. E. A. Karuza, A. E. Kahn, D. S. Bassett, *Complexity* **2019** (2019).

363. K. K. Kedzior, L. Gierke, H. M. Gellersen, M. T. Berlim, *J. Psychiatr. Res.* **75**, 107–115 (2016).

364. D. Kempe, J. M. Kleinberg, É. Tardos, *Theory of Computing* **11**, 105–147 (2015).

365. D. Kempe, J. Kleinberg, É. Tardos, presented at the SIGKDD, pp. 137–146.

366. A. N. Khambhati, A. E. Sizemore, R. F. Betzel, D. S. Bassett, *Neuroimage* **S1053-8119**, 30500–1 (2017).

367. J. Z. Kim, J. M. Soffer, A. E. Kahn, J. M. Vettel, F. Pasqualetti, D. S. Bassett, *Nat. Phys.* **14**, 91–98 (2018).

368. R. Kindermann, J. Snell, *Markov random fields and their applications* (AMS, Providence, RI, 1980).

369. O. Kinouchi, M. Copelli, *Nature Physics* **2**, 348 EP – (Apr. 2006).

370. C Kirst, M Timme, D Battaglia, *Nat Commun* **7**, 11061 (2016).

371. K Kishimoto, S.-i. Amari, *J. Math. Biol.* **7**, 303–318 (1979).

372. G. R. Kiss, C. Armstrong, R. Milroy, J. Piper, in *The computer and literary studies* (Edinburgh University Press, 1973), pp. 153–165.

373. M Kivel, A Arenas, M Barthelemy, J. P. Gleeson, Y Moreno, M. A. Porter, *J. Complex Netw.* **2**, 203–271 (2014).

374. S. C. Kleene, "Representation of events in nerve nets and finite automata," tech. rep. (RAND PROJECT AIR FORCE SANTA MONICA CA, 1951).

375. J. M. Kleinberg, *Nature* **406**, 845 (2000).

376. I. Klickstein, A. Shirin, F. Sorrentino, *Nat. Commun.* **8**, 15145 (2017).

377. I. Klickstein, A. Shirin, F. Sorrentino, *Phys. Rev. Let.* **119**, 268301 (2017).

378. C Koch, T Poggio, *Proc R Soc Lond B Biol Sci* **218**, 455–477 (1983).

379. E. Koechlin, A. Hyafil, *Science* **318**, 594–598 (2007).

380. G. H. Koenderink, Z. Dogic, F. Nakamura, P. M. Bendix, F. C. MacKintosh, J. H. Hartwig, T. P. Stossel, D. A. Weitz, *Proc. Natl. Acad. Sci.* **106**, 15192–15197 (2009).

381. B. Kosko, *Int. J. Man Mach. Stud.* **24**, 65–75 (1986).

382. M. L. Kringelbach, N. Jenkinson, S. L. Owen, T. Z. Aziz, *Nat. Rev. Neurosci.* **8**, 623 (2007).

383. J. Kunegis, presented at the Proceedings of the 22nd International Conference on World Wide Web, pp. 1343–1350.

384. Y. Kuramoto, *Chemical oscillations, waves, and turbulence* (Springer Science & Business Media, 2012), vol. 19.

385. Y. Kuramoto, H Araki, *Lecture notes in physics, international symposium on mathematical problems in theoretical physics*, 1975.

386. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science* **350**, 1332–1338 (2015).

387. D. R. J. Laming, *Information theory of choice-reaction times.* (Academic Press, 1968).

388. S. Lamrous, M. Taileb, presented at the CIMCA, pp. 18–18.

389. G. Lan, P. Sartori, S. Neumann, V. Sourjik, Y. Tu, *Nat. Phys.* **8**, 422 (2012).

390. V. Latora, M. Marchiori, *Phys. Rev. Lett.* **87**, 198701 (2001).

391. J. Lear, *Aristotle: The Desire to Understand* (Cambridge University Press, 1988).

392. J.-S. Lee, K.-I. Goh, B Kahng, D Kim, *Eur. Phys. J. B* **49**, 231–238 (2006).

393. S.-G. Lee, A. Neiman, S. Kim, *Phys. Rev. E* **57**, 3292 (1998).

394. A. M. Lesicko, T. S. Hristova, K. C. Maigler, D. A. Llano, *J Neurosci* **36**, 11037–11050 (2016).

395. J. Leskovec, A. Krevl, *SNAP Datasets: Stanford Large Network Dataset Collection*, `http://snap.stanford.edu/data`, 2014.

396. J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance, presented at the KDD, pp. 420–429.

397. J. Leskovec, J. Kleinberg, C. Faloutsos, *ACM Trans. Knowl. Discov. Data* **1**, 2 (2007).

398. A. Levina, J. M. Herrmann, T. Geisel, *Nat. Phys.* **3**, 857 (2007).

399. D. Levine, *Glauber Dynamics for Ising Model*, AMS Short Course, Tutorials, 2010.

400. P. Lévy, *Collective intelligence* (Plenum/Harper Collins New York, 1997).

401. M. Ley, presented at the International symposium on string processing and information retrieval, pp. 1–10.

402. T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, N. V. Fedoroff, *Proc. Natl. Acd. Sci.* **103**, 19033–19038 (2006).

403. A. Li, S. P. Cornelius, Y.-Y. Liu, L. Wang, A.-L. Barabási, *Science* **358**, 1042–1046 (2017).

404. X Liao, A. V. Vasilakos, Y He, *Neurosci Biobehav Rev* **77**, 286–300 (2017).

405. D. Liben-Nowell, J. Kleinberg, *Proc. Natl. Acad. Sci.* **105**, 4633–4638 (2008).

406. J. Lin, *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).

407. X. F. Liu, K. T. Chi, M. Small, *Physica A* **389**, 126–132 (2010).

408. X. Liu, J. H. Duyn, *Proc. Natl. Acad. Sci.* **110**, 4392–4397 (2013).

409. Y.-Y. Liu, A.-L. Barabási, *Rev. Mod. Phys.* **88**, 035006 (2016).

410. Y.-Y. Liu, J.-J. Slotine, A.-L. Barabási, *Nature* **473**, 167 (2011).

411. R. Lorente de Nó, *Journal für Psychologie und Neurologie* (1934).

412. A. M. Lozano, N. Lipsman, *Neuron* **77**, 406–424 (2013).

413. D. A. Luke, J. K. Harris, *Annu Rev Public Health* **28**, 69–93 (2007).

414. C. W. Lynn, D. S. Bassett, *Proc. Natl. Acd. Sci., in press* (2020).

415. C. W. Lynn, D. D. Lee, presented at the NIPS, pp. 2495–2503.

416. C. W. Lynn, D. D. Lee, *EPL* **117**, 66001 (2017).

417. C. W. Lynn, D. D. Lee, presented at the AAAI, pp. 679–686.

418. C. W. Lynn, L. Papadopoulos, D. D. Lee, D. S. Bassett, *Phys. Rev. X* **9**, 011022 (2019).

419. C. W. Lynn, A. E. Kahn, N. Nyema, D. S. Bassett, *Nat. Commun., in press* (2020).

420. C. W. Lynn, L. Papadopoulos, A. E. Kahn, D. S. Bassett, *Nat. Phys., in press* (2020).

421. D. M. MacKay, W. S. McCulloch, *Bull. Math. Biophys.* **14**, 127–135 (1952).

422. J. C. Magee, D. Johnston, *Science* **275**, 209–213 (1997).

423. R. D. Malmgren, D. B. Stouffer, A. E. Motter, L. A. Amaral, *Proc. Natl. Acd. Sci.* **105**, 18153–18158 (2008).

424. R. N. Mantegna, H. E. Stanley, *Introduction to econophysics: correlations and complexity in finance* (Cambridge University Press, 1999).

425. N. T. Markov *et al.*, *Cereb Cortex* **24**, 17–36 (2014).

426. H Markram *et al.*, *Cell* **163**, 456–492 (2015).

427. H. Markram, *Nat. Rev. Neurosci.* **7**, 153 (2006).

428. O. Marre, S. El Boustani, Y. Frégnac, A. Destexhe, *Phys. Rev. Lett.* **102**, 138101 (2009).

429. T. Martin, B. Ball, B. Karrer, M. Newman, *Phys. Rev. E* **88**, 012814 (2013).

430. M. Mäs, A. Flache, D. Helbing, *PLoS Comput Biol* **6** (2010).

431. S. Maslov, K. Sneppen, *Science* **296**, 910–913 (2002).

432. A. P. Masucci, A. Kalampokis, V. M. Eguíluz, E. Hernández-García, *PloS One* **6**, e17333 (2011).

433. M. J. Mataric, presented at the SAB, pp. 432–441.

434. G. McCarthy, E. Donchin, *Science* **211**, 77–80 (1981).

435. W. S. McCulloch, W Pitts, *Bull Math Biol* **5**, 115–133 (1943).

436. C. C. McIntyre, M. Savasta, L. Kerkerian-Le Goff, J. L. Vitek, *Clin. Neurophysiol.* **115**, 1239–1248 (2004).

437. R. McKelvey, T. Palfrey, *GEB* **7**, 6–38 (1995).

438. J. D. Medaglia, M. E. Lynall, D. S. Bassett, *J Cogn Neurosci* **27**, 1471–1491 (2015).

439. J. D. Medaglia, W Huang, E. A. Karuza, A Kelkar, S. L. Thompson-Schill, A Ribeiro, D. S. Bassett, *Nature Human Behaviour* **2**, 156–164 (2018).

440. J. D. Medaglia, D. Y. Harvey, N White, A Kelkar, J Zimmerman, D. S. Bassett, R. H. Hamilton, *J Neurosci* **38**, 6399–6410 (2018).

441. P. Mehta, D. J. Schwab, *Proc. Natl. Acad. Sci.* **109**, 17978–17982 (2012).

442. J. F. Mejias, J. D. Murray, H Kennedy, X. J. Wang, *Sci Adv.* **2**, e1601335 (2016).

443. R. S. Menon, S.-G. Kim, *Trends Cogn. Sci.* **3**, 207–216 (1999).

444. F. Meyniel, S. Dehaene, *Proc. Natl. Acad. Sci.* 201615773 (2017).

445. F. Meyniel, M. Maheu, S. Dehaene, *PLOS Comput. Biol.* **12**, e1005260 (2016).

446. M. Mézard, G. Parisi, M. Virasoro, *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications* (World Scientific Publishing Co Inc, 1987), vol. 9.

447. R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, *Science* **303**, 1538–1542 (2004).

448. G. Miritello, R. Lara, M. Cebrian, E. Moro, *Sci. Rep.* **3**, 1950 (2013).

449. B Misic, O Sporns, *Curr Opin Neurobiol* **40**, 1–7 (2016).

450. M. Molloy, B. Reed, *Random Struct. Algor.* **6**, 161–180 (1995).

451. I. Momennejad, E. M. Russek, J. H. Cheong, M. M. Botvinick, N. Daw, S. J. Gershman, *Nat. Hum. Behav.* **1**, 680 (2017).

452. I. Momennejad, A. Duker, A. Coman, *Nat. Commun.* **10**, 1–8 (2019).

453. P. R. Montague, P. Dayan, T. J. Sejnowski, *J. Neurosci.* **16**, 1936–1947 (1996).

454. A. Montanari, A. Saberi, *PNAS* **107** (2010).

455. J. Moody, *Soc. Netw.* **23**, 261–283 (2001).

456. S. A. Moosavi, A Montakhab, *Phys Rev E Stat Nonlin Soft Matter Phys* **92**, 052804 (2015).

457. T. Mora, A. M. Walczak, W. Bialek, C. G. Callan, *PNAS* **107**, 5405–5410 (2010).

458. F. Morone, H. A. Makse, *Nature* **524**, 65 (2015).

459. E. Mossel, S. Roch, presented at the STOC'07, pp. 128–134.

460. A. E. Motter, *Chaos* **25**, 097621 (2015).

461. A. E. Motter, A. P. De Moura, Y.-C. Lai, P. Dasgupta, *Phys. Rev. E* **65**, 065102 (2002).

462. M. Moussaïd, J. Kämmer, P. Analytis, H. Neth, *PLoS One* **8** (2013).

463. W. A. Mozart, *Piano Sonata No. 11*, 1784.

464. S. F. Muldoon, F Pasqualetti, S Gu, M Cieslak, S. T. Grafton, J. M. Vettel, D. S. Bassett, *PLoS Comput Biol* **12**, e1005076 (2016).

465. B. B. Murdock Jr, *J. Exp. Psychol.* **64**, 482 (1962).

466. J. Nagumo, S. Arimoto, S. Yoshizawa, *Proc. IRE* **50**, 2061–2070 (1962).

467. A. Namatame, S. Kurihara, H. Nakashima, *Emergent Intelligence of Networked Agents* (Springer, 2007), vol. 56.

468. T. Nattermann, *Spin glasses and random fields* **12**, 277 (1997).

469. H. Navarro, G. Miritello, A. Canales, E. Moro, *EPJ Data Science* **6**, 31 (2017).

470. I. Nemenman, W. Bialek, R. d. R. van Steveninck, *Phys. Rev. E* **69**, 056111 (2004).

471. Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87.

472. J. Neter, M. H. Kutner, C. J. Nachtsheim, W. Wasserman, *Applied linear statistical models* (Irwin Chicago, 1996), vol. 4.

473. C. M. Newman, *Neural Netw.* **1**, 223–238 (1988).

474. M. E. J. Newman, A Clauset, *Nature Communications* **7**, 11863 (2016).

475. M. Newman, *Networks: An Introduction* (Oxford University Press, 2010).

476. M. Newman, G. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, New York, 1999).

477. M. E. Newman, *Proc. Natl. Acad. Sci.* **98**, 404–409 (2001).

478. M. E. Newman, *SIAM Rev.* **45**, 167–256 (2003).

479. M. E. Newman, *Proc. Natl. Acad. Sci.* **103**, 8577–8582 (2006).

480. E. L. Newport, R. N. Aslin, *Cogn. Psychol.* **48**, 127–162 (2004).

481. V. Nicosia, P. E. Vértes, W. R. Schafer, V. Latora, E. T. Bullmore, *Proceedings of the National Academy of Sciences USA* **110**, 7880–7885 (2013).

482. E. Niebur, P. Erdös, *Math. Biosci.* **118**, 51–82 (1993).

483. H. Nishimori, *Statistical physics of spin glasses and information processing: An introuction* (Clarendon Press, 2001), vol. 111.

484. H. Nishimori, K. M. Wong, *Phys. Rev. E* **60**, 132 (1999).

485. A. C. Nobre, F. van Ede, *Nat. Rev. Neurosci.* **19**, 34 (2018).

486. K Norberg, B. Siejö, *Brain Res.* **86**, 45–54 (1975).

487. V. L. O'Day, R. Jeffries, presented at the Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pp. 438–445.

488. S. W. Oh, J. A. Harris, L. Ng, B. Winslow, N. Cain, S. Mihalas, Q. Wang, C. Lau, L. Kuan, A. M. Henry, *et al.*, *Nature* **508**, 207–214 (2014).

489. H Okano, A Miyawaki, K Kasai, *Philos Trans R Soc Lond B Biol Sci* **370**, 20140310 (2015).

490. J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, A.-L. Barabási, *Proc. Natl. Acd. Sci.* **104**, 7332–7336 (2007).

491. L. Onsager, *Phys. Rev.* **65**, 117 (1944).

492. M. Opper, D. Saad, *Advanced mean field methods: Theory and practice* (MIT press, 2001).

493. P. A. Ortega, D. A. Braun, *Proc. R. Soc. A* **469**, 20120683 (2013).

494. P. A. Ortega, A. A. Stocker, presented at the Adv. Neural Inf. Process. Syst. Pp. 100–108.

495. A. M. Owen, J. J. Downes, B. J. Sahakian, C. E. Polkey, T. W. Robbins, *Neuropsychologia* **28**, 1021–1034 (1990).

496. F. Pachet, P. Roy, G. Barbieri, presented at the Twenty-Second International Joint Conference on Artificial Intelligence.

497. K. Pakdaman, M. Thieullen, G. Wainrib, *Adv. Appl. Probab.* **42**, 761–794 (2010).

498. R. Palais, *Fixed Point Theory Appl.* **2**, 221–223 (2007).

499. G. Palla, I. J. Farkas, P. Pollner, I. Derényi, T. Vicsek, *New J. Phys.* **9**, 186 (2007).

500. A Palmigiano, T Geisel, F Wolf, D Battaglia, *Nat Neurosci* **20**, 1014–1022 (2017).

501. J. M. Palva, A. Zhigalov, J. Hirvonen, O. Korhonen, K. Linkenkaer-Hansen, S. Palva, *Proc. Natl. Acad. Sci.* **110**, 3585–3590 (2013).

502. L. Paninski, *Neural Comput.* **15**, 1191–1253 (2003).

503. B. Panizza, *Osservazioni sul nervo ottico* (Bernardoni, 1855).

504. P. Panzarasa, T. Opsahl, K. M. Carley, *J. Assoc. Inf. Sci. Technol.* **60**, 911–932 (2009).

505. L Papadopoulos, M. A. Porter, K. E. Daniels, D. S. Bassett, *Journal of Complex Networks* **6**, 485—565 (2018).

506. A. Paranjape, A. R. Benson, J. Leskovec, presented at the Proc. ACM WSDM, pp. 601–610.

507. H.-J. Park, K. Friston, *Science* **342**, 1238411 (2013).

508. C Parkinson, S Liu, T Wheatley, *J Neurosci* **34**, 1979–1987 (2014).

509. C Parkinson, A. M. Kleinbaum, T Wheatley, *Nat Commun* **9**, 332 (2018).

510. A. Pascual-Leone, J. R. Gates, A. Dhuna, *Neurology* **41**, 697–702 (1991).

511. A. Pascual-Leone, J. Grafman, M. Hallett, *Science* **263**, 1287–1289 (1994).

512. F Pasqualetti, S Zampieri, F Bullo, *IEEE Trans. Control Network Syst.* **1**, 40–52 (2014).

513. R. Pastor-Satorras, A. Vespignani, *Phys. Rev. Lett.* **86**, 3200 (2001).

514. C. Peng, X. Jin, K.-C. Wong, M. Shi, P. Liò, *PloS one* **7**, e34487 (2012).

515. W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images* (Elsevier, 2011).

516. M. Perc, J. J. Jordan, D. G. Rand, Z. Wang, S. Boccaletti, A. Szolnoki, *Phys. Rep.* **687**, 1–51 (2017).

517. A. E. Pereda, *Nat Rev Neurosci* **15**, 250–263 (2014).

518. J. S. Perlmutter, J. W. Mink, *Annu. Rev. Neurosci.* **29**, 229–257 (2006).

519. A. S. Persichetti, G. K. Aguirre, S. L. Thompson-Schill, *J Cogn Neurosci* **27**, 893–901 (2015).

520. S. E. Petersen, O Sporns, *Neuron* **88**, 207–219 (2015).

521. K. M. Petersson, T. E. Nichols, J.-B. Poline, A. P. Holmes, *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **354**, 1239–1260 (1999).

522. K. M. Petersson, T. E. Nichols, J.-B. Poline, A. P. Holmes, *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* **354**, 1261–1281 (1999).

523. S. J. Phillips, R. P. Anderson, R. E. Schapire, *Ecol. Modell.* **190**, 231–259 (2006).

524. S. T. Piantadosi, J. B. Tenenbaum, N. D. Goodman, *Cognition* **123**, 199–217 (2012).

525. S Piazza, P Bianchini, C Sheppard, A Diaspro, M Duocastella, *J Biophotonics* **11** (2018).

526. C Pierpaoli, P Jezzard, P. J. Basser, A Barnett, G Di Chiro, *Radiology* **201**, 637–648 (1996).

527. D. J. Pinto, G. B. Ermentrout, *SIAM J Appl Math* **62**, 206–225 (2001).

528. R. Plant, M Kim, *Biophys. J.* **16**, 227–244 (1976).

529. T. Plefka, *J. Phys. A* **15** (1982).

530. D. B. Plewes, W Kucharczyk, *J Magn Reson Imaging* **35**, 1038–1054 (2012).

531. H. Poincare, *Science and Hypothesis* (London: Walter Scott Publishing, 1905).

532. M. Polettini, M. Esposito, *Phys. Rev. Lett.* **119**, 240601 (2017).

533. M. M. Poo, J. L. Du, N. Y. Ip, Z. Q. Xiong, B Xu, T Tan, *Neuron* **92**, 591–596 (2016).

534. A. Pope, T. A. Buckley, *et al.*, *The Iliad of Homer* (WW Gibbings, 1891).

535. G Popkin, *Physicists, the Brain is Calling You*, 2016.

536. M. A. Porter, J.-P. Onnela, P. J. Mucha, *Notices of the AMS* **56**, 1082–1097 (2009).

537. J. Portugali, *The construction of cognitive maps* (Springer Science & Business Media, 1996), vol. 32.

538. M. I. Posner, C. R. Snyder, R Solso, *Cogn. Psychol.* **205** (2004).

539. J. D. Power, K. A. Barnes, A. Z. Snyder, B. L. Schlaggar, S. E. Petersen, *Neuroimage* **59**, 2142–2154 (2012).

540. R. Prakash, O. Yizhar, B. Grewe, C. Ramakrishnan, N. Wang, I. Goshen, A. M. Packer, D. S. Peterka, R. Yuste, M. J. Schnitzer, *et al.*, *Nat. Methods* **9**, 1171 (2012).

541. Queen, *Bohemian Rhapsody*, 1975.

542. J. G. Raaijmakers, R. M. Shiffrin, *Psychol. Rev.* **88**, 93 (1981).

543. M. E. Raichle, *Proc. Natl. Acad. Sci.* **95**, 765–772 (1998).

544. M. E. Raichle, *Proc Natl Acad Sci U S A* **95**, 765–772 (1998).

545. K Rajan, C. D. Harvey, D. W. Tank, *Neuron* **90**, 128–142 (2016).

546. G. Rasch, presented at the Int. Congress Psych. Vol. 2, p. 2.

547. E. Ravasz, A.-L. Barabási, *Phys. Rev. E* **67**, 026112 (2003).

548. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabási, *Science* **297**, 1551–1555 (2002).

549. N. Raz, U. Lindenberger, K. M. Rodrigue, K. M. Kennedy, D. Head, A. Williamson, C. Dahle, D. Gerstorf, J. D. Acker, *Cereb. Cortex* **15**, 1676–1689 (2005).

550. A. S. Reber, *J. Verbal Learning Verbal Behav.* **6**, 855–863 (1967).

551. F. Reif, *Fundamentals of statistical and thermal physics* (Waveland Press, 2009).

552. M. W. Reimann, M Nolte, M Scolamiero, K Turner, R Perin, G Chindemi, P Dłotko, R Levi, K Hess, H Markram, *Front Comput Neurosci* **11**, 48 (2017).

553. J. M. Reitz, *Online dictionary for library and information science* (Libraries Unlimited, 2010).

554. J. R. Reynolds, J. M. Zacks, T. S. Braver, *Cogn. Sci.* **31**, 613–643 (2007).

555. B. A. Richards, P. W. Frankland, *Neuron* **94**, 1071–1084 (2017).

556. M. Richardson, P. Domingos, *KDD'02. ACM*, 61–70 (2002).

557. J Richiardi *et al.*, *Science* **348**, 1241–1244 (2015).

558. J. P. Rickgauer, K. Deisseroth, D. W. Tank, *Nat. Neurosci.* **17**, 1816 (2014).

559. F. Rieke, D. Warland, *Spikes: Exploring the neural code* (MIT press, 1999).

560. F. Rieke, D. Warland, R. de Ruyter van Steveninck, W. Bialek, *Spikes: exploring the neural code* (MIT Press, 1997).

561. P. M. Roget, *Roget's Thesaurus of English Words and Phrases* (TY Crowell Company, 1911).

562. É. Roldán, J. M. Parrondo, *Phys. Rev. Lett.* **105**, 150607 (2010).

563. É. Roldán, J. M. Parrondo, *Phys. Rev. E* **85**, 031129 (2012).

564. A. R. Romberg, J. R. Saffran, *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 906–914 (2010).

565. R Romero-Garcia *et al.*, *Neuroimage* **171**, 256–267 (2018).

566. R Rosenbaum, M. A. Smith, A Kohn, J. E. Rubin, B Doiron, *Nat Neurosci* **20**, 107–114 (2017).

567. S. M. Ross, J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, V. L. Bristow, *Stochastic Processes* (Wiley New York, 1996), vol. 2.

568. R. A. Rossi, N. K. Ahmed, presented at the Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.

569. M. Rosvall, C. T. Bergstrom, *Proc. Natl. Acad. Sci.* **104**, 7327–7331 (2007).

570. M. Rosvall, C. T. Bergstrom, *Proc. Natl. Acad. Sci.* **105**, 1118–1123 (2008).

571. M. Rosvall, A. Trusina, P. Minnhagen, K. Sneppen, *Phys. Rev. Lett.* **94**, 028701 (2005).

572. M Rubinov, R. Ypma, C Watson, E. Bullmore, *Proceedings of the National Academy of Sciences of the USA*, doi/10.1073/pnas.1420315112 (2015).

573. E. M. Russek, I. Momennejad, M. M. Botvinick, S. J. Gershman, N. D. Daw, *PLoS Comput. Biol.* **13**, e1005768 (2017).

574. D. Rybski, S. V. Buldyrev, S. Havlin, F. Liljeros, H. A. Makse, *Proc. Natl. Acd. Sci.* **106**, 12640–12645 (2009).

575. A. J. Sadovsky, J. N. MacLean, *J Neurosci* **33**, 14048–14060 (2013).

576. J. R. Saffran, R. N. Aslin, E. L. Newport, *Science* **274**, 1926–1928 (1996).

577. R. Salvador, J. Suckling, M. R. Coleman, J. D. Pickard, D. Menon, E. Bullmore, *Cerebral cortex* **15**, 1332–1342 (2005).

578. S. Santaniello, G. Fiengo, L. Glielmo, W. M. Grill, *IEEE Trans. Neural. Syst. Rehabil. Eng.* **19**, 15–24 (2011).

579. S. Santaniello, M. M. McCarthy, E. B. Montgomery, J. T. Gale, N. Kopell, S. V. Sarma, *Proceedings of the National Academy of Sciences* **112**, E586–E595 (2015).

580. L. K. Saul, T. Jaakkola, M. I. Jordan, *Journal of artificial intelligence research* **4**, 61–76 (1996).

581. A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, B. T. Yeo, *Cereb. Cortex* **28**, 3095–3114 (2018).

582. R. Schall, *Biometrika* **78**, 719–727 (1991).

583. A Schapiro, N. Turk-Browne, presented at the Brain Mapping: An Encyclopedic Reference, pp. 501–506.

584. A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, M. M. Botvinick, *Nat. Neurosci.* **16**, 486–492 (2013).

585. T. Schelling, *Micromotives and macrobehavior* (WW Norton & Company, 2006).

586. S. J. Schiff, *Neural control engineering: the emerging intersection between control theory and neuroscience* (MIT Press, 2012).

587. R Schmalzle, M Brook O'Donnell, J. O. Garcia, C. N. Cascio, J Bayer, D. S. Bassett, J. M. Vettel, E. B. Falk, *Proc Natl Acad Sci U S A* **114**, 5153–5158 (2017).

588. J. Schmidhuber, *Neural Netw.* **61**, 85–117 (2015).

589. E. Schneidman, M. J. Berry II, R. Segev, W. Bialek, *Nature* **440**, 1007–1012 (2006).

590. A. Schnitzler, J. Gross, *Nat. Rev. Neurosci.* **6**, 285 (2005).

591. L. H. Scholtens, R Schmidt, M. A. de Reus, M. P. van den Heuvel, *J Neurosci* **34**, 12192–12205 (2014).

592. J Scholz, M. C. Klein, T. E. Behrens, H Johansen-Berg, *Nat Neurosci* **12**, 1370–1371 (2009).

593. T. Schreiber, *Phys. Rev. Lett.* **85**, 461 (2000).

594. E. Schrödinger, *What is life? The physical aspect of the living cell and mind* (Cambridge University Press Cambridge, 1944).

595. S. Schulz-Hardt, D. Frey, C. Lüthgens, S. Moscovici, *J. Pers. Soc. Psychol.* **78**, 655 (2000).

596. A Scott, *Neurophysics* (Wiley, 1977).

597. G. A. Seber, A. J. Lee, *Linear regression analysis* (John Wiley & Sons, 2012), vol. 329.

598. U. Seifert, *Phys. Rev. Lett.* **95**, 040602 (2005).

599. U. Seifert, *Rep. Prog. Phys.* **75**, 126001 (2012).

600. J. P. Sethna, *Statistical Mechanics: Entropy, Order Parameters and Complexity* (Oxford University Press, 2006).

601. H. S. Seung, U Sumbul, *Neuron* **83**, 1262–1272 (2014).

602. W. Shakespeare, *The complete works of William Shakespeare* (Wordsworth Editions, 2007).

603. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379–423 (1948).

604. C. E. Shannon, *Bell Syst. Tech. J.* **27**, 379–423 (1948).

605. G. M. Shepherd, *Foundations of the neuron doctrine* (Oxford University Press, 2015).

606. C. S. Sherrington, *The Integrative Action of the Nervous System* (Yale University Press, 1906).

607. D. Sherrington, S. Kirkpatrick, *Phys. Rev. Lett.* **35**, 1792 (1975).

608. W. L. Shew, H. Yang, T. Petermann, R. Roy, D. Plenz, *Journal of Neuroscience* **29**, 15595–15600, ISSN: 0270-6474 (2009).

609. C. T. Shih, O Sporns, S. L. Yuan, T. S. Su, Y. J. Lin, C. C. Chuang, T. Y. Wang, C. C. Lo, R. J. Greenspan, A. S. Chiang, *Curr Biol* **25**, 1249–1258 (2015).

610. J. M. Shine, M. Breakspear, P. T. Bell, K. A. E. Martens, R. Shine, O. Koyejo, O. Sporns, R. A. Poldrack, *Nat. Neurosci.* **22**, 289 (2019).

611. N. Shiraishi, T. Sagawa, *Phys. Rev. E* **91**, 012130 (2015).

612. J. S. Siegel, A. Mitra, T. O. Laumann, B. A. Seitzman, M. Raichle, M. Corbetta, A. Z. Snyder, *Cereb. Cortex* **27**, 4492–4502 (2017).

613. M. Sigman, G. A. Cecchi, *Proc. Natl. Acd. Sci.* **99**, 1742–1747 (2002).

614. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, *et al.*, *Nature* **529**, 484 (2016).

615. R. Sinatra, J. Gómez-Gardenes, R. Lambiotte, V. Nicosia, V. Latora, *Phys. Rev. E* **83**, 030103 (2011).

616. A. E. Sizemore, C. Giusti, A. Kahn, J. M. Vettel, R. F. Betzel, D. S. Bassett, *J. Comput. Neurosci.* **44**, 115–145 (2018).

617. A. E. Sizemore, E. A. Karuza, C. Giusti, D. S. Bassett, *Nat. Hum. Behav.* **2**, 682 (2018).

618. Y Sohn, M. K. Choi, Y. Y. Ahn, J Lee, J Jeong, *PLoS Comput Biol* **7**, e1001139 (2011).

619. S. Song, K. D. Miller, L. F. Abbott, *Nat. Neurosci.* **3**, 919 (2000).

620. D. Sornette, F. Deschâtres, T. Gilbert, Y. Ageon, *Phys. Rev. Lett.* **93**, 228701 (2004).

621. O Sporns, *Nat Neurosci* **17**, 652–660 (2014).

622. O Sporns, G Tononi, G. M. Edelman, *Neural Netw* **13**, 909–922 (2000).

623. O Sporns, D. R. Chialvo, M Kaiser, C. C. Hilgetag, *Trends Cogn Sci* **8**, 418–425 (2004).

624. O. Sporns, R. F. Betzel, *Annu Rev Psychol* **67**, 613–640 (2016).

625. O. Sporns, J. D. Zwi, *Neuroinformatics* **2**, 145–162 (2004).

626. O. Sporns, G. Tononi, G. M. Edelman, *Cereb. cortex* **10**, 127–141 (2000).

627. O. Sporns, G. Tononi, R. Kötter, *PLoS Comput Biol* **1**, e42 (2005).

628. K. C. Squires, C. Wickens, N. K. Squires, E. Donchin, *Science* **193**, 1142–1146 (1976).

629. K. L. Stachenfeld, M. M. Botvinick, S. J. Gershman, *Nat. Neurosci.* **20**, 1643 (2017).

630. C. J. Stam, *Nat Rev Neurosci* **15**, 683–695 (2014).

631. R. R. Stein, D. S. Marks, C. Sander, *PLoS Comp. Bio.* **11** (2015).

632. V. J. Stenger, J. A. Lindsay, P Boghossian, *Scientific American* (2015).

633. K. E. Stephan, L Kamper, A Bozkurt, G. A. Burns, M. P. Young, R Kotter, *Philos Trans R Soc Lond B Biol Sci* **356**, 1159–1186 (2001).

634. M. A. Stephens, *J. Am. Stat. Assoc.* **69**, 730–737 (1974).

635. S. Sternberg, *Am. Sci.* **57**, 421–457 (1969).

636. M. Steyvers, J. B. Tenenbaum, *Cogn. Sci.* **29**, 41–78 (2005).

637. J Stiso, D. S. Bassett, *arXiv* **1807**, 04691 (2018).

638. C. Stosiek, O. Garaschuk, K. Holthoff, A. Konnerth, *Proc Natl Acad Sci U S A* **100**, 7319–7324 (2003).

639. S. H. Strogatz, *Nature* **410**, 268 (2001).

640. S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, W. Bialek, *Phys. Rev. Lett.* **80**, 197 (1998).

641. B. Stuhrmann, M. S. e Silva, M. Depken, F. C. MacKintosh, G. H. Koenderink, *Phys. Rev. E* **86**, 020901 (2012).

642. M. P. Stumpf, M. A. Porter, *Science* **335**, 665–666 (2012).

643. L. Šubelj, M. Bajec, presented at the Proceedings of the 22nd international conference on World Wide Web, pp. 527–530.

644. H. J. Sussmann, V. Jurdjevic, *Differ. Equ.* **12**, 95–116 (1972).

645. R. S. Sutton, in *Machine Learning Proceedings 1995* (Elsevier, 1995), pp. 531–539.

646. R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning* (MIT press Cambridge, 1998), vol. 2.

647. G. S. Sylvester, *J. Stat. Phys.* **33**, 91–98 (1983).

648. K. Sznajd-Weron, J. Sznajd, *Int. Mod. Phys. C* **11**.

649. S. Y. Takemura *et al.*, *Nature* **500**, 175–181 (2013).

650. T. Tanaka, *Phys, Rev. E* **58**, 2302 (1998).

651. A. Tang, D. Jackson, J. Hobbs, W. Chen, J. L. Smith, H. Patel, A. Prieto, D. Petrusca, M. I. Grivich, A. Sher, *et al.*, *J. Neurosci.* **28**, 505–518 (2008).

652. E. Tang, D. S. Bassett, *Rev. Mod. Phys.* **90**, 031003 (2018).

653. E. Tang, C. Giusti, G. L. Baum, S. Gu, E. Pollock, A. E. Kahn, D. R. Roalf, T. M. Moore, K. Ruparel, R. C. Gur, *et al.*, *Nat. Commun.* **8**, 1252 (2017).

654. P Tass, M. Rosenblum, J Weule, J Kurths, A Pikovsky, J Volkmann, A Schnitzler, H.-J. Freund, *Phys. Rev. Lett.* **81**, 3291 (1998).

655. P. N. Taylor, J Thomas, N Sinha, J Dauwels, M Kaiser, T Thesen, J Ruths, *Front Neurosci* **9**, 202 (2015).

656. P. N. Taylor, Y Wang, M Kaiser, *Sci Rep* **7**, 39859 (2017).

657. M. Teboulle, *First Order Algorithms for Convex Minimization*, IPAM, Tutorials, 2010.

658. J. B. Tenenbaum, T. L. Griffiths, C. Kemp, *Trends Cogn. Sci.* **10**, 309–318 (2006).

659. T Teşileanu, B Olveczky, V Balasubramanian, *Elife* **6**, e20944 (2017).

660. The Beatles, *A Hard Day's Night*, 1964.

661. *SIAM REVIEW* **45**, 167—256 (2003).

662. B. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, *et al.*, *J. Neurophysiol.* **106**, 1125–1165 (2011).

663. P. M. Thompson, T. D. Cannon, K. L. Narr, T. Van Erp, V.-P. Poutanen, M. Huttunen, J. Lönnqvist, C.-G. Standertskjöld-Nordenstam, J. Kaprio, M. Khaledy, *et al.*, *Nat. Neurosci.* **4**, 1253 (2001).

664. D. J. Thouless, P. W. Anderson, R. G. Palmer, *Philosophical Magazine* **35**, 593–601 (1977).

665. G. Tkačik, O. Marre, D. Amodei, E. Schneidman, W. Bialek, M. J. Berry II, *PLoS Comput. Biol.* **10**, e1003408 (2014).

666. E. C. Tolman, *Psychol. Rev.* **55**, 189 (1948).

667. S. H. Tompson, A. E. Kahn, E. B. Falk, J. M. Vettel, D. S. Bassett, *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 253 (2019).

668. G Tononi, M Boly, M Massimini, C Koch, *Nat Rev Neurosci* **17**, 450–461 (2016).

669. G. Tononi, O. Sporns, G. M. Edelman, *Proc. Natl. Acad. Sci.* **91**, 5033–5037 (1994).

670. Toto, *Africa*, 1982.

671. J. Travers, S. Milgram, *Phychology Today* **1**, 61–67 (1967).

672. F. Tria, V. Loreto, V. D. P. Servedio, S. H. Strogatz, *Sci. Rep.* **4**, 5890 (2014).

673. N. B. Turk-Browne, P. J. Isola, B. J. Scholl, T. A. Treat, *J. Exp. Psychol. Learn. Mem. Cogn.* **34**, 399 (2008).

674. A. Tversky, D. Kahneman, *Science* **185**, 1124–1131 (1974).

675. W. J. Tyler, *Nature Reviews Neuroscience* **13**, 867–878 (2012).

676. L. Van Aelst, C. D'Souza-Schorey, *Genes Dev.* **11**, 2295–2322 (1997).

677. M. P. Van Den Heuvel, H. E. H. Pol, *Eur Neuropsychopharmacol* **20**, 519–534 (2010).

678. D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E Yacoub, K Ugurbil, W.-M. H. Consortium, *Neuroimage* **80**, 62–79 (2013).

679. M. P. van den Heuvel, O Sporns, *J Neurosci* **31**, 15775–15786 (2011).

680. M. P. van den Heuvel, E. T. Bullmore, O Sporns, *Trends Cogn Sci* **20**, 345–361 (2016).

681. M. P. van den Heuvel, O. Sporns, *Trends Cogn. Sci.* **17**, 683–696 (2013).

682. R. d. R. van Steveninck, W. Bialek, *Proc. R. Soc. Lond. B* **234**, 379–414 (1988).

683. A. Vázquez, J. G. Oliveira, Z. Dezsö, K.-I. Goh, I. Kondor, A.-L. Barabási, *Phys. Rev. E* **73**, 036127 (2006).

684. P. E. Vertes, A. F. Alexander-Bloch, N Gogtay, J. N. Giedd, J. L. Rapoport, E. T. Bullmore, *Proc Natl Acad Sci U S A* **109**, 5868–5873 (2012).

685. R. Vicente, M. Wibral, M. Lindner, G. Pipa, *J. Comput. Neurosci.* **30**, 45–67 (2011).

686. W. E. Vinje, J. L. Gallant, *Science* **287**, 1273–1276 (2000).

687. B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, presented at the Proceedings of the 2nd ACM workshop on Online social networks, pp. 37–42.

688. M. S. Vitevitch, K. Y. Chan, S. Roodenrys, *J. Mem. Lang.* **67**, 30–44 (2012).

689. H. von Helmholtz, *Archiv für Anatomie, Physiologie und wissenschaftliche Medicin*, 71–73 (1850).

690. J von Neumann, *The Computer and the Brain*.

691. S. Vossel, J. J. Geng, G. R. Fink, *Neuroscientist* **20**, 150–159 (2014).

692. V Vuksanovic, P Hovel, *Neuroimage* **97**, 1–8 (2014).

693. E Wallace, H. R. Maei, P. E. Latham, *Neural Comput* **25**, 1408–1439 (2013).

694. V. Walsh, A. Cowey, *Nat. Rev. Neurosci.* **1**, 73 (2000).

695. V. Walsh, A. Ellison, L. Battelli, A. Cowey, *Proc. R. Soc. Lond., B, Biol. Sci.* **265**, 537–543 (1998).

696. X. Wang, A. McCallum, presented at the SIGKDD, pp. 424–433.

697. Z. Wang, C. T. Bauch, S. Bhattacharyya, A. d'Onofrio, P. Manfredi, M. Perc, N. Perra, M. Salathé, D. Zhao, *Phys. Rep.* **664**, 1–113 (2016).

698. L. M. Ward, *Trends Cogn. Sci.* **7**, 553–559 (2003).

699. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440–442 (1998).

700. V. J. Wedeen, P. Hagmann, W.-Y. I. Tseng, T. G. Reese, R. M. Weisskoff, *Magn. Reson. Med.* **54**, 1377–1386 (2005).

701. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, *Proc. Natl. Acd. Sci.* **106**, 67–72 (2009).

702. R. West, J. Leskovec, presented at the Proceedings of the 21st international conference on World Wide Web, pp. 619–628.

703. A. J. Whalen, S. N. Brennan, T. D. Sauer, S. J. Schiff, *Phys. Rev. X* **5**, 011005 (2015).

704. K. J. Whitaker *et al.*, *Proc Natl Acad Sci U S A* **113**, 9105–9110 (2016).

705. J. G. White, E Southgate, J. N. Thomson, S Brenner, *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).

706. S. D. Whitehead, L.-J. Lin, *Artif. Intell.* **73**, 271–306 (1995).

707. W. Whitman, *Leaves of grass* (Oregan Publishing, 2017).

708. W. A. Wickelgren, *Acta Psychol.* **41**, 67–85 (1977).

709. R. Wilensky (1983).

710. H. R. Wilson, J. D. Cowan, *Biophys. J.* **12**, 1–24 (1972).

711. E. Winograd, D. S. Lynn, *Mem. Cogn.* **7**, 29–34 (1979).

712. J. M. Wolfe, T. S. Horowitz, N. M. Kenner, *Nature* **435**, 439 (2005).

713. W. L. Woodrow, W. P. Clawson, J Pobst, Y Karimipanah, N. C. Wright, R Wessel, *Nature Physics* **11**, 659–663 (2015).

714. S. R. y Cajal, *Estructura de los centros nerviosos de las aves*.

715. G Yan, P. E. Vertes, E. K. Towlson, Y. L. Chew, D. S. Walker, W. R. Schafer, A. L. Barabasi, *Nature* **550**, 519–523 (2017).

716. M. Yates, *J. Exp. Psychol. Learn. Mem. Cogn.* **39**, 1649 (2013).

717. J. Yedidia, *Advanced mean field methods: Theory and practice*, pp. 459–468.

718. H. Yin, I. Artsimovitch, R. Landick, J. Gelles, *Proc. Natl. Acad. Sci.* **96**, 13124–13129 (1999).

719. M. P. Young, J. W. Scannell, G. A. Burns, C Blakemore, *Rev Neurosci* **5**, 227–250 (1994).

720. Q Yue, R. C. Martin, S Fischer-Baum, A. I. Ramos-Nunez, F Ye, M. W. Deem, *J Cogn Neurosci* **29**, 1532–1546 (2017).

721. W. W. Zachary, *J. Anthropol. Res.* **33**, 452–473 (1977).

722. A Zalesky, A Fornito, E Bullmore, *Neuroimage* **60**, 2096–2106 (2012).

723. G. Zamora-López, Y. Chen, G. Deco, M. L. Kringelbach, C. Zhou, *Sci. Rep.* **6**, 38424 (2016).

724. E. Zarahn, G. K. Aguirre, M. D'Esposito, *Neuroimage* **5**, 179–197 (1997).

725. S. Zeki, *A vision of the brain* (Blackwell Scientific Publ., 1993).

726. H. L. Zeng, E. Aurell, M. Alava, H. Mahmoudi, *Physical Review E* **83**, 041135 (2011).

727. X Zhang, C Moore, M. E. J. Newman, *Eur. Phys. J. B* **90**, 200 (2017).

728. M Zhen, A. D. Samuel, *Curr Opin Neurobiol* **33**, 117–126 (2015).

729. R. Zia, B Schmittmann, *J. Stat. Mech.* **2007**, P07012 (2007).

730. X. N. Zuo, Y He, R. F. Betzel, S Colcombe, O Sporns, M. P. Milham, *Trends Cogn Sci* **21**, 32–45 (2017).

731. J Zylberberg, A Pouget, P. E. Latham, E Shea-Brown, *PLoS Comput Biol* **13**, e1005497 (2017).